DOCUMENT RESUME

ED 439 143 TM 030 678

AUTHOR Lee, Guemin
TITLE Conditional Standard Errors of Measurement for Tests Composed of Testlets.
PUB DATE 1999-04-20
NOTE 66p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Definitions; *Error of Measurement; Estimation (Mathematics); *Reliability; Statistical Bias; Test Items
IDENTIFIERS *Testlets

ABSTRACT

Previous studies have indicated that the reliability of test scores composed of testlets is overestimated by conventional item-based reliability estimation methods (S. Sireci, D. Thissen, and H. Wainer, 1991; H. Wainer, 1995; H. Wainer and D. Thissen, 1996; G. Lee and D. Frisbie). In light of these studies, it seems reasonable to ask whether the item-based estimation methods for the conditional standard errors of measurement (SEM) would provide underestimates for tests composed of testlets. The primary purpose of this study was to investigate the appropriateness and implication of incorporating a testlet definition into the estimation procedures of the conditional SEM for tests composed of testlets. Another purpose was to investigate the bias in estimates of the conditional SEM when using item-based methods instead of testlet-based methods. Several estimation procedures were proposed and compared in estimating conditional SEM for tests composed of testlets. Conditions under which these might be used are described. (Contains 4 tables, 23 figures, and 40 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 439 143

Conditional Standard Errors of Measurement for Tests Composed of Testlets

Guemin Lee
CTB/McGraw-Hill

BEST COPY AVAILABLE

Paper Presented at the 1999 Annual Meeting of the National Council on Measurement in Education
Montreal, Canada
April 20 1999

TM030678

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Guemin Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Abstract

Previous studies have indicated that the reliability of test scores composed of testlets is overestimated by conventional item-based reliability estimation methods (Sireci, Thissen & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Lee & Frisbie, in press). In light of these previous studies, it seems reasonable to ask whether the item-based estimation methods for the conditional standard error of measurement (SEM) would provide underestimates for tests composed of testlets. The primary purpose of this study was to investigate the appropriateness and implication of incorporating a testlet definition into the estimation procedures of the conditional SEM for tests composed of testlets. Another purpose was to investigate the bias in estimates of the conditional SEM when using item-based methods instead of testlet-based methods. Several estimation procedures were proposed and compared in estimating conditional SEM for tests composed of testlets.

Conditional Standard Errors of Measurement for Tests Composed of Testlets

Testlets, as the name implies, have been defined as small tests (Wainer & Kiely, 1987; Wainer & Lewis, 1990). Previous studies have indicated that the reliability of test scores composed of testlets is overestimated by conventional item-based reliability estimation methods (Sireci, Thissen & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Lee & Frisbie, in press). That is, when subgroups of items in a test are related to the same passage or other stimulus material, there might be statistical dependence among those items, causing an item-based reliability estimate to be inflated relative to an estimate of reliability based on the correlation between equivalent forms (Lawrence, 1995). In light of these previous studies, it seems reasonable to ask whether the item-based estimation methods for the conditional standard error of measurement (conditional SEM) would provide underestimates for tests composed of testlets. This question was the main motivation for doing this study.

When measurement models are applied in practical situations, some statistical assumptions must be made, such as conditional independence (or uncorrelated errors) and unidimensionality. Because the unidimensional measurement models based on dichotomously scored items are frequently used for practical applications, it is important to study the robustness of these models to violation of their assumptions in various applied contexts. Previous studies have shown that the assumptions for measurement are frequently violated by tests composed of testlets (Sireci, Thissen & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Lee & Frisbie, in press; Lee, Kolen, Frisbie & Ankenmann, 1998). Therefore, applying unidimensional measurement models based on dichotomously scored items to estimating conditional SEM for tests composed of testlets might be inappropriate. Because there is little evidence in the literature about how the violation of assumptions affects estimates of conditional SEM, it is not clear how serious the degree of distortion of the conditional SEM estimates might be.

The primary purpose of this study was to investigate the appropriateness and implication of incorporating a testlet definition into the estimation procedures of the conditional SEM for tests composed of testlets. This study also investigated the bias in the estimates of the conditional SEM based

on using item-based methods instead of testlet-based methods when the assumptions required by measurement modeling have been violated.

The objectives of this study were:

1. To investigate the relative appropriateness of each of several methods by making a comparison between the prespecified true conditional SEM and the estimates obtained from each method.

2. To assess the relative magnitude of bias introduced by using each method in estimating the conditional SEM for tests composed of testlets.

3. To examine the robustness of the item-based methods with respect to violation of the conditional independence assumption in estimating the conditional SEM for tests composed of testlets

4. To investigate the relationship between the degree of violation of the conditional independence assumption and the degree of bias in estimates of the conditional SEM.

Methods of Estimating Conditional SEM

In classical test theory, the standard error of measurement is estimated by $\hat{\sigma}_E = S_X\sqrt{1-\hat{\rho}_{XX'}}$, where S_X is the standard deviation of a set of test scores and $\hat{\rho}_{XX'}$ is the reliability estimate for those test scores. This formula, which can be viewed as an average standard error of measurement, provides one estimate for all examinees, regardless of their score level (Qualls-Payne, 1992). However, it is reasonable to expect that the amount of error associated with individual scores could vary depending on where the true score is located on the score scale.

Since the first edition of the Test Standards, the American Psychological Association, American Educational Research Association and National Council on Measurement in Education (1954), have recommended that test publishers estimate and report the standard error of measurement at several points on the score scale. The current version, Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985), also included this recommendation in Standard 2.10 (p.22).

A number of methods have been developed to estimate the conditional SEM. The earliest investigators about the conditional SEM were probably Mollenkopf (1949) and Thorndike (1951). Lord

(1955, 1957) developed the best-known conditional SEM estimation formula using binomial error theory. Feldt (1984) provided another estimation method using a compound binomial error model, which presumes that parallel forms involve stratified random samples of items. An item response theory (IRT) approach to estimating the conditional SEM was provided by Lord (1980), and recently a generalizability theory (G-theory) approach was presented by Brennan (1998). These methods can be thought of as the fundamental frameworks for estimating conditional SEMs, and several variations of these basic frameworks may be possible. A comprehensive review of most of these and related methods is summarized in Feldt & Brennan (1989) and Feldt & Qualls (1996).

However, the issues related to estimating the conditional SEM for tests composed of testlets have not been addressed so far. (Brennan, 1998, investigated this issue under a generalizability theory framework, however, he did not mention the testlet concept explicitly.) The estimation methods for the conditional SEM were classified in this study as either item-based or testlet-based. The IRT and G-theory approaches were considered for estimating the conditional SEM for each item-based and testlet-based method. Because Lord's binomial error model (1955, 1957) and Feldt's compound binomial error model (1984) are special cases of the G-theory approach for estimating the conditional SEM (Brennan, 1998), the IRT and G-theory approaches together include almost all basic formulas mentioned above, except variations from Thorndike's (1951) and Mollenkopf's (1949) methods.

Two item-based estimation methods were considered: (a) A G-theory approach with a pxI design [pxI method], where p represents persons, the object of measurement, and I represents the item facet, and (b) a dichotomous IRT approach [DIRT method]. A G-theory approach with a px(I:H) design [px(I:H) method], where p represents persons, H represents the passage facet, and I represents the item facet within a passage, and polytomous IRT approaches for estimating conditional SEM using both Samejima's (1969) graded response model [GIRT method] and Bock's (1972) nominal model [NIRT method] were used as the testlet-based estimation methods.

Item-Based Methods

These methods, which assume that the appropriate measurement unit is an item, have been used most frequently for estimating the conditional SEM. Here, it is assumed that items are scored dichotomously, although the underlying methodology per se makes no such assumption (Brennan, 1998).

pxI Method

The pxi G-theory design is appropriate for estimating the conditional SEM where i represents an item facet composed of an infinite, undifferentiated set of items, and p represents an object of measurement, a person in this case. Typically, it is assumed that the objects of measurement "facet" is infinite. Let X_{pi} denote an observed score for person p on item i. Then, the X_{pi} can be represented as:

$$
\begin{aligned}
X_{pi} &= \mu && \text{(grand mean)} \\
&+ \mu_p - \mu && \text{(person effect)} \\
&+ \mu_i - \mu && \text{(item effect)} \\
&+ X_{pi} - \mu_p - \mu_i + \mu && \text{(residual effect).}
\end{aligned}
\qquad [1]
$$

In this linear model for a pxi G-study design, the decomposition in Equation [1] is for single person–item combinations. Therefore, estimated variance components from a G-study are also for single items. However, decisions are to be based on a total (or mean) score for a set of items. The linear model for such a mean score is based on a pxI D-study design, and a linear model for a D-study design is the same as in Equation [1], except for replacing i with I in all terms containing i. So, the variance components in a D-study are for a set of items and not for a single item.

Two types of decisions can be differentiated in the G-theory framework: relative and absolute decisions. Corresponding to these two types of decisions, two types of errors can also be differentiated: relative and absolute errors (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991; Brennan, 1992). In this study, only absolute errors are considered in comparing various estimation

methods because the most other methods are based only on the absolute error definition. The absolute conditional SEM for person p can be estimated by

$$\hat{\sigma}(\Delta)_p = \sqrt{\frac{\sum (X_{pi} - X_{pI})^2}{I'(I-1)}} \qquad [2]$$

where X_{pI} is person *p*'s mean score over I items, I is the number of items in the G-Study, and I' is the number of items in the D-Study (Brennan, 1998).

DIRT Method

This method is based on an item response curve, representing the probability that individual person *k* with ability score θ_k will answer item i correctly, denoted $P_i(\theta_k)$. In this study, the three parameter logistic model was used for obtaining the item response curve. To estimate the conditional SEM using an IRT approach, it is necessary to obtain the distribution of the number-correct raw scores given IRT ability (θ) with estimated item parameters (Kolen, Zeng & Hanson, 1996). The probability of random variable X representing a certain raw score on a K-item test for ability θ can be denoted as $P(X = i|\theta)$, where i ranges from 0 to K. This notation expresses the conditional distribution of the number-correct raw scores for a given ability level. Lord & Wingersky (1984) provided a recursion formula to calculate these probabilities:

$$\begin{aligned} P(X = i|\theta) &= P(X_{r-1} = i|\theta)[1 - P(\theta)] && i = 0 \\ &= P(X_{r-1} = i|\theta)[1 - P(\theta)] + P(X_{r-1} = i-1|\theta)P(\theta) && 0 < I < r \\ &= P(X_{r-1} = i-1|\theta)P(\theta) && i = r. \end{aligned} \qquad [3]$$

The variance of the resulting distribution is the conditional error variance of the number-correct raw scores for ability θ. Therefore, the conditional SEM for a given θ can be estimated by taking the square root of this conditional error variance (Kolen, Zeng & Hanson, 1996).

Testlet-Based Methods

The testlet concept has been recommended as a useful tool for solving the problems arising from the situations in which the conditional independence assumption among items is violated. (Thissen, Steinberg & Mooney, 1989; Sireci, Thissen & Wainer, 1991; Wainer, Sireci & Thissen, 1991; Yen, 1993; Wainer, 1995; Wainer & Thissen, 1996, Lee, Kolen, Frisbie & Ankenmann, 1998). The polytomous IRT approaches incorporate this recommendation. The G-theory approach, however, can take the passage (or testlet) facet into account as another source of variation (Lee & Frisbie, in press).

NIRT Method

With respect to testlet applications, Bock's nominal model has been used predominantly (Wainer & Thissen, 1996; Sireci, Thissen, & Wainer, 1991; Wainer, Sireci, & Thissen, 1991) because "the testlet scores are nominal (or at most semi-ordered) responses; as we show later, a score of 1 may not always reflect higher proficiency than a score of 0, due to guessing" (Thissen, Steinberg, & Mooney, 1989). This could be the reason that Bock's nominal model has been used in this situation: polytomous models other than Bock's nominal model assume ordered response categories.

Under Bock's (1972) nominal model, the probability that an examinee with a given ability (θ) responds to category k in passage j is

$$P_{jk}(\theta) = \frac{\exp[a_{jk}\theta + c_{jk}]}{\sum_{k=1}^{K} \exp[a_{jk}\theta + c_{jk}]}, \qquad [4]$$

where j=1,2,....J (passages), k=1,2,....,K (categories). The constraints, $\sum_k a_{jk} = \sum_k c_{jk} = 0$, are imposed on this model. The parameters of this model are rescaled by using centered polynomials of the associated scores to represent the category-to-category changes in the a_k and c_k values: $a_{jk} = \sum_{p=1}^{P} \alpha_{jp}(k - \frac{K}{2})^p$ and $c_{jk} = \sum_{p=1}^{P} \gamma_{jp}(k - \frac{K}{2})^p$, where the parameters, $[\alpha_p, \gamma_p]j$, p = 1,2, ..., P for $p \le K$, are the free parameters to be estimated from the data (Thissen, Steinberg, & Mooney, 1989).

The next procedures for estimating the conditional SEM is similar to the application of the dichotomous IRT models. For this procedure, it is necessary to obtain the distribution of number-correct raw scores given IRT ability (θ) under a polytomous model. Hanson (1994) extended the Lord & Wingersky (1984) algorithm to polytomous items. The recursive algorithm is (Wang, Kolen & Harris, 1996):

For item i, [5]

$$P_1(X = x|\theta) = P(U_1 = x|\theta) \qquad x = 0, 1, 2, \ldots, n_1$$

For item k=2,3,4,...,K,

$$P_k(X = x|\theta) = \sum_{u=0}^{n_k} P_{k-1}(X = x - u)P(U_k = u|\theta) \qquad x = 0, 1, 2, \ldots, \sum_{k=1}^{K} n_k$$

In Equation [5], the U_k represents a random variable for the score on item *k* with scores from 0 to n_k. The appropriate probabilities can be obtained from Equation [4] . The variance of the resulting distribution is the conditional error variance of the number-correct raw scores for ability θ , and the conditional SEM for a given θ can be estimated by taking the square root of this conditional error variance.

GIRT Method

In this study, Samejima's (1969) graded response model was used, as well as Bock's (1972) nominal model, in order to check on the possibility of using polytomous IRT models based on ordered categories. Samejima's (1969) graded response model seems appropriate for estimating conditional SEMs. There would be an ordered quality to testlet-based scores if such scores corresponded to the extent of completeness of the examinee's reasoning process within a specific testlet. This seems to be a reasonable representation for reading comprehension testlets, where several dichotomous items relate to a single reading passage. The more of such items within a testlet that an examinee answers correctly, the more extensive is his or her comprehension. Therefore, in the present study, Samejima's (1969) graded response model was compared to Bock's (1972) nominal model with respect to performance in estimating the conditional SEM for tests composed of testlets (Lee, Kolen, Frisbie & Ankenmann, 1998).

Under Samejima's (1969) graded response model, consider passage *j*, in which the number-correct score corresponding to the dichotomous items that constitute the passage can be classified into one of *K* categories, numbered 1 through *K* inclusive with consecutive integers, and "call such a response a 'graded response'..." (p.20). Then, the probability that a graded response to passage *j* is classified into category *k* or higher, given θ, is

$$P_{jk}^{*}(\theta) = \begin{cases} 1 & k = 1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} & 2 \le k \le K \\ 0 & k > K \end{cases} \qquad [6]$$

The parameter a_j is the passage discrimination parameter, which is constant across the response categories of a particular passage (i.e., constant throughout the whole reasoning process). The $b_{j,k-1}$ is the difficulty parameter of the category boundary *k*-1 ($2 \le k \le K$) for passage *j*, and it is free to vary among the category boundaries of a particular passage such that $b_{j,k-1} < b_{j,k}$. (Note that $b_{j,k-1}$ is the θ-value at which the probability of the response being classified into category *k* or higher is 0.5.) The probability that a graded response is classified in category *k*, given θ, is defined by $P_{jk}(\theta) = P_{jk}^{*}(\theta) - P_{j,k+1}^{*}(\theta)$, which is also written as

$$P_{jk}(\theta) = \begin{cases} 1 - \dfrac{1}{1+\exp[-a_j(\theta - b_{j1})]} & k = 1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} - \dfrac{1}{1+\exp[-a_j(\theta - b_{jk})]} & 2 \le k \le K-1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} & k = K \end{cases} \qquad [7]$$

The examinee's number-correct score distribution can be obtained by using Equations [7] and [5]. Then, the conditional SEM for a given θ can be estimated by following the same procedures that are used in the NIRT method.

px(I:H) Method

The univariate px(i:h) design, persons (p) crossed with items (i) nested in passages (h), is appropriate for estimating the conditional SEM for this situation. The linear model for the response of a person to an item within a passage treats persons as objects of measurement and items and passages as random facets. For this model, n_p persons represent a random sample from a population of interest, and n_h passages represent a random sample from the universe of passages. The $n_{i:h}$ items in a passage are also considered a random sample from the universe of items related to that passage. This linear model, referred to as completely random, can be represented as:

$$
\begin{aligned}
X_{pih} = {} & \mu && \text{(grand mean)} \\
& + \mu_p - \mu && \text{(person effect)} \\
& + \mu_h - \mu && \text{(passage effect)} \\
& + \mu_{i:h} - \mu_h && \text{(item within passage effect)} \\
& + \mu_{ph} - \mu_p - \mu_h + \mu && \text{(person by passage interaction effect)} \\
& + X_{pih} - \mu_{ph} - \mu_{i:h} + \mu_h && \text{(residual effect)}
\end{aligned}
\qquad [8]
$$

where p=1, ... ,n_p ; i=1, ... ,$n_{i:h}$; and h=1, ... ,n_h .

A linear model for a D-study design is the same as in Equation [8], except for replacing i and h with I and H, respectively, in all terms containing i and h. Then, as Brennan (1996) has shown, the absolute conditional SEM can be computed by

$$
\sigma(\Delta)_p = \sqrt{\frac{\sigma^2(h)_p}{H} + \frac{\sigma^2(i{:}h)_p}{I_+}} \text{, where } I_+ = \sum_h I_h \text{ .} \qquad [9]
$$

where H represents the number of passages (or testlets) in the D-study. The I_h represents the number of items within the *h*th passage in the D-study.

Simulations

Model for simulations

Response data sets of tests composed of stimulus-based testlets (e.g., Reading Comprehension tests) were simulated. Nandakumar (1991) provided a method of simulating a paragraph comprehension test data set. According to her method, *k* items of a paragraph comprehension test are split into *h* groups of items. Two abilities are considered to have influence on the examinee's response to each item: one is common to all items of the test (denoted as θ_g in this paper) and the other is unique to each group (denoted as θ_h, h=1, 2, 3, ... , H, where H represents the number of passages in this paper). That is, the examinee's response to a certain item is influenced by general ability (θ_g) and passage-specific ability (θ_h). For example, if there are H passages in a test, H+1 ("1" represents a general ability influencing all items in the test) abilities would be considered. These H+1 abilities are assumed to be independent, standard normal random variables. She also introduced a bivariate extension of the unidimensional three-parameter logistic model with compensatory abilities:

$$P_i(\theta_g,\theta_h) = c_i + \frac{1 - c_i}{1 + \exp\{-1.7[a_{gi}(\theta_g - b_{gi}) + a_{hi}(\theta_h - b_{hi})]\}}, \qquad [10]$$

where $P_i(\theta_g,\theta_h)$ is the probability that an examinee having θ_g and θ_h ability scores answers item *i* correctly,

a_{gi} and a_{hi} are the discrimination parameters of item *i* for general and passage–specific ability dimensions, respectively,

b_{gi} and b_{hi} are the difficulty parameters of item *i* for general and passage-specific ability dimensions, respectively, and

c_i is the lower asymptote parameter of item *i*.

For simulating the data set for this study, the parameters shown in Equation [10] need to be selected. The item difficulty parameters b_{gi} and b_{hi} were taken from independent, identical normal

distributions. The item discrimination parameters a_{gi} and a_{hi} were generated using the following equations from Nandakumar (1991):

$$a_{gi} \sim N\{(1-\xi)\mu, \sqrt{1-\xi}\,\sigma\} \qquad [11]$$
$$a_{hi} \sim N\{\xi\mu, \sqrt{\xi}\,\sigma\}$$
$$a_{gi} + a_{hi} \sim N(\mu, \sigma)$$

where μ and σ represent the mean and the standard deviation, respectively, of the discrimination parameter for a test. The ξ can be interpreted as the degree of influence of each passage-specific ability relative to the general ability on an item. For example, if ξ is equal to zero, then the examinee's response depends upon only the level of general ability. As the value of the ξ increases, the influence of the passage-specific ability increases. Consequently, the conditional dependence among items within passages would increase. In this way, it is possible to manipulate the level of conditional dependence among items within passages by specifying different values of ξ.

Procedures for simulating data sets

The model for simulations discussed so far is based on a two-dimensional IRT approach with compensatory abilities. In this model, passage-specific abilities were considered as one factor influencing an examinee's response, with general ability being another factor. The conceptualization of this model treats passages as a fixed facet, not a random one. However, this conceptualization is different from the one adopted in this paper. Previously in this paper, the passage facet was considered a random facet. In order to incorporate this different conceptualization about the test into the Nandakumar (1991) procedures, the data were simulated as follows:

Step 1. Specify a test composed of testlets.

1-1 Fix the total number of items, k (e.g., k=42).

1-2. Split k items into h groups of items (e.g., h_1=6, h_2=6, h_3=6, h_4=6, h_5=6, h_6=6, h_7=6).

Step 2. Generate a population of persons based on general ability. Select n examinees randomly from the general ability scale, θ_g, assuming θ_g is distributed as standard normal (e.g., n=1000).

Step 3. Specify a test form.

Generate the item parameters from the distributions defined in Equation [11].

Step 4. Generate passage specific abilities for each examinee of the generated population for a specified test form.

For each selected examinee, generate passage-specific abilities on the scale θ_h, h=1, 2, 3, ... , H, assuming each θ_h being independently distributed standard normal.

Step 5. Generate a response data set.

5-1. Compute the probability of a correct answer to item i for each examinee using Equation [10]. Then, create a matrix A, which is composed of n rows (representing examinees) by k columns (representing items) using computed probabilities.

5-2. Generate random numbers from a uniform distribution U(0, 1) and create a matrix B, which consists of elements with the dimensions of n x k.

5-3. From a comparison of elements between matrices A and B, generate matrix C, composed of 0 or 1. Assign 1 to c_{ij}, if b_{ij} is equal to or less than a_{ij}, and otherwise, assign 0 to c_{ij}.

From these steps, an examinee's response data set consisting of 0 and 1 can be obtained. Repeating the procedures from step 3 to step 5 would make another examinee's response data set. In these procedures, the general ability of each examinee was fixed (not included in the repeated loop), and the passage specific abilities of each examinee were selected from the specified distributions (included in the repeated loop). These procedures can be thought of as a modification of the procedures used by Nandakumar (1991), permitting the passages to be considered a random facet, not fixed. That is, the examinee's passage-specific abilities were assumed to change across randomly sampled passages. These data simulation procedures were required for obtaining the true conditional SEM, which was used as a criterion for comparing various estimation methods for tests composed of testlets.

Preliminary Analyses for Simulations

The purposes of doing the preliminary analyses for sets of simulations were: (1) to make the simulated data sets as similar as possible to the real data sets and (2) to determine the appropriate ξ values. The need for doing the second preliminary analysis relates to the two dimensional compensatory model that was used in this study.

The decision between the compensatory and non-compensatory models is a somewhat subjective one. It seems reasonable to apply the compensatory model to the testlet situations rather than applying the non-compensatory model within the appropriate range of ξ values. For example, suppose that a certain student takes a reading comprehension test and the first passage deals with the topic of baseball games. Also, suppose the value of ξ is in a reasonable range (e.g., about 0.3). If that student's reading comprehension ability level (general ability in this study) is in the middle score range but passage–specific ability for the first passage (baseball) is in the high score range, then the probability that the student answers items associated with the first passage would be expected to be slightly higher compared to the probability when considering reading comprehension ability (general ability) alone. That is, student's high passage–specific ability could compensate his/her lower general ability on answering a given item correctly to yield a score somewhat above the middle range.

However, the compensatory model has some limitations. For example, assume the same test situations and a relatively high ξ value (e.g., about 0.7). Even though the student's reading comprehension ability is extremely low (e.g., θ_g=-3.0), if the passage–specific ability is extremely high (e.g., θ_h=3.0), a very high probability of correctly answering items belonging to that passage would be expected. This case seems somewhat unreasonable and unrealistic. Therefore, even though the compensatory model could be more reasonable than the non-compensatory model, it should be used under a reasonable range of ξ values. Checking this range was the second reason for doing the preliminary analyses.

In explaining simulation procedures earlier, the procedures for selecting μ and σ for Equation [11] were not described thoroughly, even though it was mentioned that these values are the mean and standard deviation of the discrimination parameter. Nandakumar (1991) selected these values from real

data sources such as the SAT verbal test battery, the ACT mathematics test battery, the Armed Services Vocational Aptitude Battery for auto shop information, and so forth. So for this study, the means, standard deviations, and maximum and minimum values of the discrimination, difficulty, and lower asymptote parameter estimates based on three-parameter logistic model for several Iowa Tests of Basic Skills (ITBS) tests composed of testlets are reported in Table 1.

Insert Table 1 About Here

The means and standard deviations of item parameter estimates for the grade 7 Reading Comprehension test were initially selected as inputs to Equations [10] and [11] for the first step of preliminary analyses. (The c parameter in Equation [10] was fixed to 0.2 for all simulated items.) Then, the simulation procedures were applied under each ξ value specified in Table 2, and the degree of dependence measure and general characteristics of simulated data sets are presented in the same table.

Insert Table 2 About Here

The ξ values ranged from 0.1 to 0.6, with an interval of 0.1. For Step 1 in Table 2, the means of Yen's (1984) Q_3 statistics for between- and within-passage item pairs are most similar to those of the target (between-passage: -0.022, within-passage: 0.027) for the ξ value of 0.5. When the values of ξ are less than 0.5, the means of between-passage Q_3 statistics are similar to each other, but the means of within-passage Q_3 statistics are different from those of the target with the different ξ values. The positive relationship between the mean of within-passage Q_3 statistics and the ξ value can be found in this table. This result seems to be reasonable, because this positive relationship between conditional dependence and the ξ value could be explained by the logic embedded in the simulation model used in this study.

However, one important finding can be observed by examining the general characteristics between the target and the six simulated data sets. That is, even though the means of the Q_3 statistics for between- and within-passage item pairs for the ξ value of 0.5 are similar to those of the target, the mean discrimination parameter of the simulated data set is much smaller than that of the target.

Furthermore, the ... tendency for the mean discrimination parameter estimates to shrink more, compared to that o... .e target, as the value of ξ increases. In contrast, the mean of the difficulty parameter has a much greater value compared to that of the target. The mean of the lower asymptote parameter estimates is slightly higher than that of the target. In sum, the general characteristics of the item parameter estimates under the ξ value of 0.5 are very different from those of the target.

Another important check would be to compare the mean and standard deviation of the target and simulated data sets. Using the mean and standard deviation of proportion–correct scores would be more sensible than using raw scores because the grade 7 Reading Comprehension test and the simulated data sets have different total numbers of items. From this comparison, non-negligible differences can also be observed. In short, the item parameter estimates and general characteristics of the simulated data set under the ξ value of 0.5 are too different from those of the target, even though the conditional dependence measures from both data sources are similar.

Based on these results, inputs to Equations [10] and [11] for simulations were changed by using a linear estimation. For example, in Step 1, the μ and σ in Equation [11] were assumed as 0.952 and 0.287 which were derived from Table 1, but in Step 2, they were modified to be 1.630 and 0.553, respectively, by setting linear equations of (0.952:0.556 = ? : 0.952) and (0.287:0.149 = ? :0.287). The other parameter specifications for simulations were computed by using the same linear estimation method. The simulation procedures were applied with the new set of parameter specifications. The general characteristics of the simulated data sets were computed and are presented in Table 2 under the heading of Step 2.

In contrast to the results from Step 1, the means of the Q_3 statistics for between- and within-passage item pairs under the ξ value of 0.3 are most similar to those of the target. Descriptive statistics about the item parameter estimates and general characteristics of simulated data set are much more similar to those of the target than are those from Step 1. Therefore, in this study, parameter specifications under Step 2 were used for the subsequent simulation studies.

The next issue is associated with selecting appropriate specific ξ values. Based on the results presented in Table 2, it might be reasonable to set ξ values around 0.3 for simulations. The ξ values

from 0.2 to 0.4 with an interval of 0.025 (ξ values of 0.200, 0.225, 0.250,..., 0.350, 0.375, and 0.400) were examined by investigating the conditional dependence measures and preliminary results by applying various conditional SEM estimation methods. As a result, 0.275, 0.300, 0.325, and 0.350 were selected for the ξ values for simulating data responses. The relationship between the specified ξ values and conditional dependence measures are presented in Table 3.

Insert Table 3 About Here

In order to investigate the relationship between specified ξ values and the degree of conditional dependence, these measures examining the degree of conditional dependence were applied to each simulated data set under each specified value of ξ. As anticipated, the degree of conditional dependence among within-passage items increases, as the ξ value goes up. The interpretation of the results for these conditional dependence measures is the same as was given earlier for examining the conditional independence assumption for the real data sets. In general, a ξ value of 0.275 can be understood to represent somewhat mild violation of the assumptions compared to the real data sets used in this study. The ξ values of 0.300 and 0.325 provide conditional dependence measures similar to those obtained from the real data sets. These two ξ values provide for a moderate violation of the assumptions. The ξ value of 0.350 provides conditional dependence measures indicating a severe violation of the assumptions compared to the real data sets.

Criterion Indexes for Simulations

It would be informative and convenient to formulate overall indexes to represent the degree of error involved in using each estimation method. First, the error can be conceptualized as the difference between an estimate of conditional SEM for each examinee from using a particular estimation method and the true conditional SEM for that examinee:

$$\hat{sem}_{pr} - sem_p \qquad [12]$$

where sem_p is a true conditional SEM for a person p and $\hat{sem}_{pr}$ is an estimated conditional SEM for the same person p on a particular replication r. This error can be divided into two parts: bias induced by a

particular estimation method and the random error over replications. In this study, 50 replications were conducted and then these two components were disentangled as:

$$[\hat{sem}_{pr} - \overline{\hat{sem}_p}] + [\overline{\hat{sem}_p} - sem_p] \qquad [13]$$

where $\overline{\hat{sem}_p}$ is the average of the conditional SEMs for person p over replications. (In this study, the fitted mean, obtained from a polynomial regression, was used for this average value of the conditional SEMs.) The first part represents random error, and the second part represents bias associated with using a particular estimation method.

Based on the above conceptualization, three indexes were developed: average root–mean–squared error (ARMSE), average root–mean–squared bias (ARMSB), and average standard error of estimate (ASEE):

$$ARMSE = \sqrt{\frac{1}{PR}\sum_p \sum_r (\hat{sem}_{pr} - sem_p)^2}$$

$$ARMSB = \sqrt{\frac{1}{P}\sum_p (\overline{\hat{sem}_p} - sem_p)^2} \qquad [14]$$

$$ASEE = \sqrt{\frac{1}{PR}\sum_p \sum_r (\hat{sem}_{pr} - \overline{\hat{sem}_p})^2}$$

where P represents the total number of simulees and R represents the total number of replications. One advantage of using these indexes is that the variance of total error can be decomposed into two parts: one for squared bias and the other for random error variance. That is, the equation $ARMSE^2 = ARMSB^2 + ASEE^2$ always holds.

Analysis Strategies

The true conditional SEM was obtained so that it could be compared with estimates using various estimation methods applied to the simulated data set. In the previous section, the data

simulation procedures were outlined in five steps. In order to get the true conditional SEM for each selected examinee, the procedures from Step 3 to Step 5 were repeated the specified number of times. For each simulated data set, the total score of each examinee was computed, and then the standard deviation for these r total scores for each examinee was computed. Each standard deviation for total test scores of each examinee can be thought of as his/her true conditional SEM, if r goes to infinity. In this study, data generation procedures were replicated 1000 times, and 1000 (assumed) true conditional SEMs for 1000 examinees selected in Step 2 were computed. These true conditional SEMs served as criteria for estimates obtained using various item-based or testlet-based estimation methods.

One more data set was generated using the same simulation procedures to obtain an examinee's response data set. Using this data set, the item-based and testlet-based conditional SEM estimation methods were applied. For the G-theory approach, a computer application program (Brennan, 1996) was used to estimate the conditional SEM for each pxI or px(I:H) design. For IRT methods, the BILOG (Mislevy & Bock, 1990) and MULTILOG (Thissen, 1991) computer programs were used for estimating item parameters and ability parameters. The number–correct raw score distribution for given theta values was formulated, and the conditional SEM was computed by a FORTRAN90 application program written for this purpose. The estimate from each method was then compared with the true conditional SEM of each examinee. These comparison procedures were repeated 50 times to control the error of estimates that may influence the magnitude of the estimated conditional SEM. From these results, the most appropriate method for estimating the conditional SEM for tests composed of testlets was determined, and also the most robust method among item-based methods was identified.

In order to investigate the relationship between conditional dependence and bias in estimates of the conditional SEM using item-based methods, the above procedures for comparing various estimation methods were repeated under certain prespecified values of ξ (0.275, 0.300, 0.325, and 0.350). To make the interpretation of ξ more meaningful, the relationship between different ξ values and level of conditional dependence was investigated. The generalizability of the results from analyzing the simulated data sets was checked with the real data sets.

Results

Results from Simulations (ξ =0.275)

Figure 1 shows comparisons between the true conditional SEM and the mean of estimated conditional SEMs over 50 replications of using each estimation method. The horizontal axis in each graph of the figure represents a true score scale, which was computed by averaging the total test scores of examinees over 1000 replications, following the steps outlined in previous section. In order to get the true conditional SEMs, the standard deviation of the total scores of each examinee over 1000 replications was computed, and a curve was fitted to the SEMs of 1000 examinees to obtain the true conditional SEM. The mean of the estimated conditional SEMs were obtained by averaging the conditional SEM estimates over 50 replications for each estimation method. That is, each method was applied to each replication and repeated 50 times.

Insert Figure 1 About Here

The pxI method provides estimates of conditional SEM that are similar to the true conditional SEM, even though it slightly overestimates the conditional SEM in the middle score range. The conditional SEM estimates of the px(I:H) method are similar to the true conditional SEM, but it also slightly overestimates conditional SEM in the middle score range. This method also has much larger fluctuations within true scores than do the other estimation methods. The DIRT method provides smaller estimates of the conditional SEM compared to the true conditional SEM. That is, the DIRT method underestimates the conditional SEM of test scores based on testlets. The GIRT and NIRT methods provide estimates of the conditional SEM that are similar to each other. In the middle score range, the estimates from these two polytomous IRT estimation methods are similar to the true conditional SEM, but in the lower and higher score ranges, they overestimate the conditional SEM.

To get more general trends, the fitted line of conditional SEM estimates of using each method are plotted in Figure 2 along with a line for the true conditional SEM. The fitted line of the conditional SEM estimates of each method was obtained by applying a polynominal regression technique. In the middle score range, all estimation methods except the DIRT method provide similar estimates of conditional

SEM. But in the lower and higher score ranges, the GIRT and NIRT methods give higher estimates. This overestimation is a little bit greater in the GIRT method compared to the NIRT method in the lower score range. The pxI and px(I:H) methods provide almost the same estimates of conditional SEM along the true score scale.

Insert Figure 2 About Here

Bias lines, based on the fitted line from each estimation method, and the true conditional SEM as a baseline are presented in Figure 3. The bias trends are similar for the pxI and px(I:H) methods. That is, both methods provide slightly positive bias in the middle score range. The DIRT estimation method gives negatively biased estimates throughout the score scale. Even though the bias lines for both polytomous IRT models seem to be more dramatic than the one from the DIRT method, the influence of bias in a practical sense would be much greater with the DIRT method compared to the polytomous IRT estimation methods. That is, because the distribution of true scores is similar to the normal distribution, the bias in the middle score range would be more severe and influential than the bias in the extremes due to the larger number of examinees affected.

Insert Figure 3 About Here

Discussion so far has focused on the bias introduced by each estimation method in terms of a fitted line and did not consider the error of estimates. Figure 4 shows the standard error of estimate of using each estimation method. Much larger standard errors of estimate can be found in the px(I:H) method compared to the other estimation methods. That is, the px(I:H) method provides fitted conditional SEM estimates that are similar to true conditional SEM, but these estimates contain relatively large amounts of error.

Insert Figure 4 About Here

Three indexes of error associated with each estimation method under four specified ξ values are presented in Table 4. Under the ξ value of 0.275, the pxI method provides the smallest ARMSE.

Therefore, if there is a need to estimate the conditional SEM for each person on one administration of a test, the pxI method would produce relative small amounts of error. (The GIRT and NIRT methods would be similar.) However, the ARMSB of the px(I:H) method is the smallest one, which means this method introduces the least amount of bias in estimating conditional SEM for each person. Even though this method has the smallest value of ARMSB, it has the biggest value of ASEE. The proportion of variance of total error explained by the error of estimate is about 99.3% for the px(I:H) estimation method.

Insert Table 4 About Here

In comparing the DIRT and polytomous IRT methods, the polytomous IRT methods provide smaller ARMSE and ARMSB values, and they provide ASEE values similar to the DIRT method. The NIRT estimation method seems to be only slightly better than the GIRT estimation method in the context of mild violation of assumptions for measurement modeling.

Results from Simulations (ξ =0.300)

Comparisons of the true conditional SEM and mean of the estimated conditional SEMs over 50 replications of using each estimation method are presented in Figure 5. The pxI method underestimates the conditional SEM in the middle of the score range. The underestimation of the DIRT method here is much more evident compared to the results from the ξ value of 0.275 in Figure 1. Both the GIRT and NIRT estimation methods provide slightly underestimated conditional SEMs in the middle score range and overestimated conditional SEMs in the lower and higher score ranges. The px(I:H) method provides estimates of conditional SEM similar to the true conditional SEM, but it has much greater error of estimate compared to the other estimation methods.

Insert Figure 5 About Here

Insert Figure 6 About Here

The fitted line of the true conditional SEM and fitted lines for the estimated conditional SEM using each method are provided in Figure 6. The pxI, NIRT, and GIRT methods provide similar estimates of the conditional SEM in the middle score range. The px(I:H) method provides the highest conditional SEM estimates in the middle score range, while in the lower and higher score ranges, the GIRT and NIRT estimation methods do.

Insert Figure 7 About Here

Figure 7 shows the bias lines from each estimation method on the true score scale. The px(I:H) method provides conditional SEM estimates that are quite similar to the true conditional SEM, even though it overestimates a little in both the lower and higher score ranges. The pxI method underestimates the conditional SEM in the middle score range (around from 13 to 32). The DIRT method underestimates the conditional SEM along almost all the score range. The NIRT method underestimates conditional SEM a little bit more in the middle score range than the GIRT method. Much larger standard errors of estimate can be identified in the px(I:H) method by comparing the standard error of estimate plots among estimation methods presented in Figure 8. These results are very similar to those shown in Figure 4 for ξ =0.275.

Insert Figure 8 About Here

According to Table 4, the pxI method provides the smallest ARMSE, but the px(I:H) method provides the smallest ARMSB. Both the GIRT and NIRT estimation methods provide much smaller ARMSE and ARMSB compared to the DIRT method. The GIRT method provides a little bit smaller ARMSE and ARMSB compared to the NIRT method, but both methods have similar ASEE values.

Results from Simulations (ξ =0.325)

Figure 9 shows comparisons between the true conditional SEM and the mean of estimated conditional SEMs using each estimation method under the ξ value of 0.325. Basically, the trends observed in this figure are similar to those found in Figure 5 from the ξ value of 0.300, except for two differences. First, the pxI method provides much smaller estimates of conditional SEM compared to the

true conditional SEM along the true score scale. Second, the discrepancy between the true conditional SEM and the estimate of conditional SEM from the DIRT method becomes greater when the ξ value moves from 0.300 to 0.325. The fitted conditional SEM for each method and the true conditional SEM are presented in Figure 10. The bias lines of the estimation methods and the standard error of estimate plots are provided in Figure 11 and Figure 12, respectively. Compared to bias lines from the ξ value of 0.300, the G-theory approaches produce different trends. The IRT approaches yield trends of bias lines that are similar to those from ξ =0.300.

Insert Figure 9 About Here

Insert Figure 10 About Here

Insert Figure 11 About Here

Insert Figure 12 About Here

According to Table 4, the px(I:H) method provides a much smaller ARMSB value compared to the other estimation methods, but it still has the largest ASEE value. Both polytomous IRT methods provide much smaller ARMSE values compared to the other estimation methods.

Results from Simulations (ξ =0.350)

The results from simulations under the ξ value of 0.350 are presented in Figures 13, 14, 15, and 16. Similar trends and interpretations can be observed and made as in investigating the results from simulations under the ξ value of 0.325. The main difference is that the degree of bias increased as the ξ value changed from 0.325 to 0.350. According to Table 4, the px(I:H) method provides much smaller ARMSB than do the other methods. The GIRT method provides the smallest ARMSE value, even though the NIRT method has the smaller ARMSB value.

Insert Figure 13 About Here

Insert Figure 14 About Here

Insert Figure 15 About Here

Insert Figure 16 About Here

Relationship between Degree of Violation of Assumptions and Bias in Estimates of the Conditional SEM

One of the research objectives of this study was to investigate the relationship between the degree of violation of the assumptions required by measurement modeling and the amount of bias in the estimates of the conditional SEM using item-based methods instead of testlet-based methods. To address this objective, bias lines for each of the four specified values of ξ are replotted in the same graph, all shown in Figure 17, for the purpose of comparison. As discussed in explaining Table 3, the ξ values have a positive relationship with the degree of conditional dependence.

Insert Figure 17 About Here

According to Figure 17, in top left graph for pxI method, bias increases as the ξ value goes up (ignoring the ξ value of 0.275). This finding can be confirmed by the overall indexes in Table 4. The ARMSB of the pxI method changes in accordance with the change of the ξ values: ARMSBs 0.108, 0.217, 0.349 and ξ values of 0.300, 0.325, 0.350, respectively. The reason for excluding the results from the ξ value of 0.275 is that the pxI method has a tendency to overestimate the conditional SEM for unidimensional tests (Agresti & Coull, 1998; Lee, Brennan & Kolen, 1998); it overestimates the conditional SEM under the situation of the ξ value of 0.275. Therefore, the results from the ξ value of 0.275 would not be appropriate for investigating bias trends here. By comparing the bias of the DIRT method for specified ξ values, it is evident that there is a positive relationship between the degree of bias and the degree of violation of assumptions. The values of ARMSB changes 0.254, 0.394, 0.463, and 0.589 with the change of the ξ values, 0.275, 0.300, 0.325, and 0.350, respectively.

Reducing the Standard Error of Estimate in the px(I:H) Method

The px(I:H) estimation method provides the smallest ARMSB and the highest ASEE for all conditions of the simulations in this study. It provides the highest ARMSE value compared to the other

estimation methods. Consequently, this method would not be a good choice for estimating the conditional SEM of each examinee on one test administration, even though it introduces the least bias. However, if it is possible to reduce the error of estimate of the px(I:H) method, then this method would have an important advantage over the other methods in estimating conditional SEMs for tests composed of testlets in practical situations. Two techniques could be considered in the practical use of this method. One is to use the fitted estimates of conditional SEMs, and the other is to report conditional SEMs at only integer score points.

Brennan (1998) indicated that considerable errors were involved in estimates from px(I:H) method and suggested that the fitted estimates be used rather than the unfitted ones. He also argued that "this seems especially appropriate when the number of observations within objects of measurement is small and the number of objects of measurement is large (p.33)." This situation seems to be applied to each replication of the simulations used in this study. The fitted estimates of conditional SEM using a quadratic function were computed for each replication for the ξ value of 0.325, and the ARMSE, ARMSB, and ASEE were calculated.

Figure 18 shows the comparison between the true conditional SEM and the mean of fitted estimates of conditional SEM (fitted px(I:H) method). The fitted px(I:H) method provides the estimates of conditional SEM similar to the true conditional SEMs. By comparing this figure with the top-right graph in Figure 9, much less variation of points can be observed. The bias line for the fitted px(I:H) method is presented in Figure 19. Based on the comparison with the top-right graph in Figure 11, a little bit larger bias can be found, which is mainly due to the overestimation compared to the true conditional SEM. The standard errors of estimate of the fitted px(I:H) method are plotted in Figure 20. Much smaller standard errors of estimate were obtained by using the fitted estimates of conditional SEMs instead of the unfitted ones, which can be confirmed by comparing this figure with the top-right graph in Figure 12. According to Table 4, the fitted px(I:H) method produces much smaller ARMSE and ASEE, but larger ARMSB values compared to those of the px(I:H) method. Even though the magnitude of ASEE for the px(I:H) method decreases by using the fitted estimates of conditional SEMs rather than the unfitted ones, it is still the highest value compared to the other methods. Because, from a practical standpoint, it would be

sensible to use the fitted estimates of conditional SEMs instead of the unfitted ones, this technique seems to be a promising one for reducing the standard error of estimate of the px(I:H) estimation method.

Insert Figure 18 About Here

Insert Figure 19 About Here

Insert Figure 20 About Here

Aggregating the conditional SEM on integer score points is another technique for reducing error of estimate. According to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985), conditional SEMs should be reported at appropriate, well-separated levels or intervals. In this study, the conditional SEM for each integer score point was recalculated by grouping examinees based on their true scores. For example, in order to get an aggregated estimate of the conditional SEM for the true score of 18, the average of the conditional SEM estimates over examinees having true scores between 17.5 and 18.5 was computed. This idea was applied to obtaining both true conditional SEMs and the estimates of the px(I:H) method on integer score points, which are reported in Figure 21. In this figure, the data sets from the ξ value of 0.325 were used. The px(I:H) method provides similar estimates of the conditional SEM on integer score points compared to the true conditional SEMs.

Insert Figure 21 About Here

Figure 22 shows the bias of the px(I:H) estimation method under these new estimations. This representation of bias is very similar to the one in Figure 11. The standard error of estimate of the px(I:H) estimation method for each integer score point is given in Figure 23. Much smaller errors of estimate are observed compared to those using the conditional SEM estimates of individual examinees, which are presented in the top-right plot in Figure 12. Comparing both plots, the errors of estimate

decrease from about 0.9 to 0.2. Because the conditional SEMs are provided for integer score points in practice by many testing companies, this technique is a promising method for estimating conditional SEM for tests composed of testlets.

Insert Figure 22 About Here

Insert Figure 23 About Here

Discussion

Based on findings of this study, these conclusions are offered:

First, in general, the item-based estimation methods, both the pxI and DIRT methods, underestimate the conditional SEM for tests composed of testlets. However, the pxI method provides good estimates of the conditional SEM under mild violation of the assumptions, and this method is more robust to the violation of the assumptions compared to the DIRT method. The robustness of the pxI estimation method might be due to its tendency to overestimate the conditional SEM for a unidimensional test.

Second, the px(I:H) method introduces the smallest amount of bias, but the largest error of estimate. This method seems to be the best estimation method for tests composed of testlets in terms of the magnitude of bias. One way to reduce the error of estimate dramatically is to use a quadratic fit, as discussed by Brennan (1998). Also, reporting conditional SEMs at well-separated score intervals seems to be an efficient way of reducing the error of estimate.

Third, the GIRT and NIRT methods provide similar estimates of the conditional SEM. Therefore, the use of Samejima's graded response model seems to be as appropriate as Bock's nominal model, at least, with respect to performance in estimating the conditional SEM for tests composed of testlets. Both methods provide estimates of the conditional SEM that are similar to the true conditional SEM in the middle score range, but they overestimate the conditional SEM in the lower and higher score ranges. This overestimation might be caused by loss of information when testlet scores are used as the unit of

analysis, as indicated by Yen (1993). These methods provide good estimates of the conditional SEMs under moderate and somewhat severe violation of assumptions.

Fourth, the bias of the item-based estimation methods increases as the degree of conditional dependence goes up. That is, an increase in the extent of violation of the assumptions required by measurement modeling leads to a corresponding increase in bias in the estimates of the conditional SEM for tests composed of testlets.

REFERENCES

Agresti, A., & Coull, B.A. (1998). Approximate is better than "Exact" for interval estimation of binomial proportions. The American Statistician, 52, 119-126.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51 (suppl.), 201-238.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Brennan, R.L. (1992). Elements of generalizability theory. Iowa City, IA: ACT.

Brennan, R.L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. Applied Psychological Measurement, 22, 307-331.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item response theory. Journal of Educational and Behavioral Statistics, 22, 265-289.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements : Theory of generalizability for scores and profiles. New York: Wiley.

Feldt, L.S. (1984). Some relationships between the binomial error model and classical test theory. Educational and Psychological Measurement, 44, 883-891.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), Educational measurement.(3rd ed.). Phoenix, AZ: Oryx Press.

Feldt, L.S., & Qualls, A.L. (1996). Estimation of measurement error variance at specific score levels. Journal of Educational Measurement, 33, 141-156.

Hanson, B.A. (1994). An extension of the Lord-Wingersky algorithm to the polytomous items. Unpublished research note.

Hoover, H.D., Hieronymus, A.N., Frisbie, D.A., & Dunbar, S.B. (1994) Iowa Tests of Basic Skills : Interpretive guide for school administrators. Chicago, IL: The Riverside Publishing Company.

Kolen, M.J., Zeng, L., & Hanson, B.A. (1996). Conditional standard errors of measurement for scale scores using IRT. Journal of Educational Measurement, 33, 129-140.

Lawrence, I.M. (1995). Estimating reliability for tests composed of item sets. (RR-95-18). Princeton, NJ: ETS.

Lee, G., & Frisbie, D.A. (in press). Estimating reliability under a generalizability theory model for test scores composed of testlets. Applied Measurement in Education.

Lee, G., Kolen, M.J., Frisbie, D.A., & Ankenmann, R.D. (1998, April). Equating test forms composed of testlets using dichotomous and polytomous IRT models. Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Lee, W., Brennan, R.L., & Kolen, M.J. (1998, April). A comparison of some procedures for estimating conditional scale-score standard errors of measurement. (Iowa Testing Programs Occasional Paper No. 43). Iowa City, IA: University of Iowa.

Lord, F.M. (1955). Estimating test reliability. Educational and Psychological Measurement, 15, 325-336.

Lord, F.M. (1957). Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 17, 510-521.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating." Applied Psychological Measurement, 8, 453-461.

Mislevy, R.J., & Bock, R.D. (1990). BILOG 3. Item analysis and test scoring with binary logistic models (2nd ed.). Mooresville, IN:Scientific Software.

Mollenkopf, W.G. (1949). Variation of the standard error of measurement. Psychometrika, 14, 189-229.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-117.

Qualls-Payne, A.L. (1992). A comparison of score level estimates of the standard error of measurement. Journal of Educational Measurement, 29, 213-225.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. Psychometric Monograph Supplement, 17.

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

Thissen, D. (1991). MULTILOG Multiple categorical item analysis and test scoring using item response theory (version 6.0). Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical models. Journal of Educational Measurement, 26, 247-260.

Thorndike, R.L. (1951). Reliability. In E.F. Lindquist (Ed.), Educational measurement. Washington, DC: American Council on Education.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. Applied Measurement in Education, 8, 157-186.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing : A case for testlets. Journal of Educational Measurement, 24, 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement : Issues and Practice, 15, 22-29.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential testlet functioning: Definition and detecting. Journal of Educational Measurement, 28, 197-219.

Wang, T., Kolen, M.J., & Harris, D.J. (1996). Conditional standard errors, reliability and decision consistency of performance levels using polytomous IRT. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Table 1
Descriptive Statistics of Item Parameter Estimates for Several ITBS Tests Composed of Testlets

	Reading Grade 4	Reading Grade 7	Maps Grade 4	Maps Grade 7
Mean a_i	0.805	0.952	0.961	0.807
S.D. a_i	0.251	0.287	0.330	0.225
Max a_i	1.449	1.748	1.673	1.343
Min a_i	0.384	0.427	0.499	0.436
Mean b_i	0.355	0.342	0.212	0.782
S.D. b_i	0.851	0.960	0.824	0.670
Max b_i	2.059	2.405	1.635	1.952
Min b_i	-1.039	-1.776	-1.534	-0.309
Mean c_i	0.163	0.202	0.175	0.194
S.D. c_i	0.036	0.052	0.055	0.045
Max c_i	0.248	0.337	0.282	0.320
Min c_i	0.090	0.127	0.094	0.141

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams, Vocab = Vocabulary, Sim = simulated.

Table 2

Characteristics of Simulated Data Sets for Specified ξ Values

Criterion	Target	ξ=0.1	ξ=0.2	ξ=0.3	ξ=0.4	ξ=0.5	ξ=0.6
		Step 1					
Mean of Q_3							
Between	-.022	-.016	-.016	-.019	-.016	-.022	-.020
Within	.027	-.018	-.006	.002	.013	.033	.034
S.D. of Q_3							
Between	.044	.033	.034	.033	.034	.033	.035
Within	.061	.031	.034	.033	.036	.038	.046
Mean	25.4	20.9	20.9	20.9	21.5	20.7	21.2
S.D.	9.08	7.78	7.83	6.73	5.93	5.51	4.83
Mean of Prop	.552	.498	.498	.498	.512	.493	.505
S.D. of Prop	.197	.185	.186	.160	.141	.131	.115
Mean of a_i's	.952	1.068	1.129	.715	.596	.556	.521
S.D. of a_i's	.287	.335	.323	.173	.186	.149	.172
Mean of b_i's	.342	.784	1.013	.771	1.018	.973	1.399
S.D. of b_i's	.960	.778	.720	1.135	1.600	1.316	1.836
Mean of c_i's	.202	.256	.289	.219	.250	.236	.280
S.D. of c_i's	.052	.061	.085	.041	.053	.030	.059
		Step 2					
Mean of Q_3							
Between	-.022	-.017	-.019	-.023	-.028	-.032	-.035
Within	.027	-.015	-.004	.024	.057	.093	.118
S.D. of Q_3							
Between	.044	.035	.036	.035	.036	.033	.035
Within	.061	.047	.038	.048	.044	.064	.063
Mean	25.4	22.5	22.4	22.3	21.9	22.55	21.7
S.D.	9.08	9.71	9.40	8.13	8.12	7.19	6.31
Mean of Prop	.552	.536	.533	.531	.521	.536	.517
S.D. of Prop	.197	.231	.224	.194	.193	.171	.150
Mean of a_i's	.952	1.291	1.111	.891	.834	.702	.626
S.D. of a_i's	.287	.355	.343	.360	.243	.209	.177
Mean of b_i's	.342	.297	.224	.486	.370	.385	.664
S.D. of b_i's	.960	.915	.958	1.003	.925	1.206	1.205
Mean of c_i's	.202	.175	.173	.199	.178	.210	.221
S.D. of c_i's	.052	.034	.034	.065	.033	.046	.044

Note. Target = graded 7 Reading Comprehension test, Mean of Prop = mean of proportion correct scores, S.D. of Prop = standard deviation of proportion correct scores.

Table 3
Descriptive Statistics for Q_3 Statistics for Four Specified ξ Values

Q_3 Statistics		ξ=0.275	ξ=0.300	ξ=0.325	ξ=0.350
Mean					
	Between	-0.021	-0.022	-0.025	-0.026
	Within	0.016	0.022	0.029	0.042
S.D.					
	Between	0.038	0.042	0.035	0.035
	Within	0.053	0.055	0.051	0.049

Table 4

Average Root Mean Squares of Error (ARMSE), Average Root Mean Square of Bias (ARMSB), and Average Standard Error of Estimate (ASEE) for Each Estimation Method for Four Values of ξ

Method	ARMSE	ARMSB	ASEE
		ξ=0.275	
pxI	.219	.096 (19.4%)	.197 (80.6%)
px(I:H)	.832	.071 (0.7%)	.829 (99.3%)
DIRT	.289	.254 (77.0%)	.139 (23.0%)
GIRT	.227	.183 (65.1%)	.134 (34.9%)
NIRT	.222	.175 (62.5%)	.136 (37.5%)
		ξ=0.300	
pxI	.237	.108 (20.7%)	.211 (79.3%)
px(I:H)	.844	.083 (1.0%)	.840 (99.0%)
DIRT	.423	.394 (86.8%)	.153 (13.2%)
GIRT	.264	.223 (71.4%)	.141 (28.6%)
NIRT	.275	.232 (71.3%)	.147 (28.7%)
		ξ=0.325	
pxI	.320	.217 (46.0%)	.235 (54.0%)
px(I:H)	.886	.040 (0.2%)	.885 (99.8%)
fitted px(I:H)	.344	.106 (9.6%)	.327 (90.4%)
DIRT	.496	.463 (87.1%)	.179 (12.9%)
GIRT	.240	.173 (52.0%)	.167 (48.0%)
NIRT	.239	.159 (44.2%)	.179 (55.8%)
		ξ=0.350	
pxI	.425	.349 (67.4%)	.243 (32.6%)
px(I:H)	.917	.065 (0.5%)	.915 (99.5%)
DIRT	.618	.589 (90.8%)	.188 (9.2%)
GIRT	.267	.195 (53.2%)	.184 (46.8%)
NIRT	.271	.182 (45.0%)	.201 (55.0%)

Note. pxI=G-theory estimation method with a pxI design, px(I:H)=G-theory estimation method with a px(I:H) design, DIRT=dichotomous IRT estimation method, GIRT=graded response model estimation method, NIRT=nominal model estimation method. The number within parenthesis represents the percentage of variation of total error explained by bias or error of estimate.

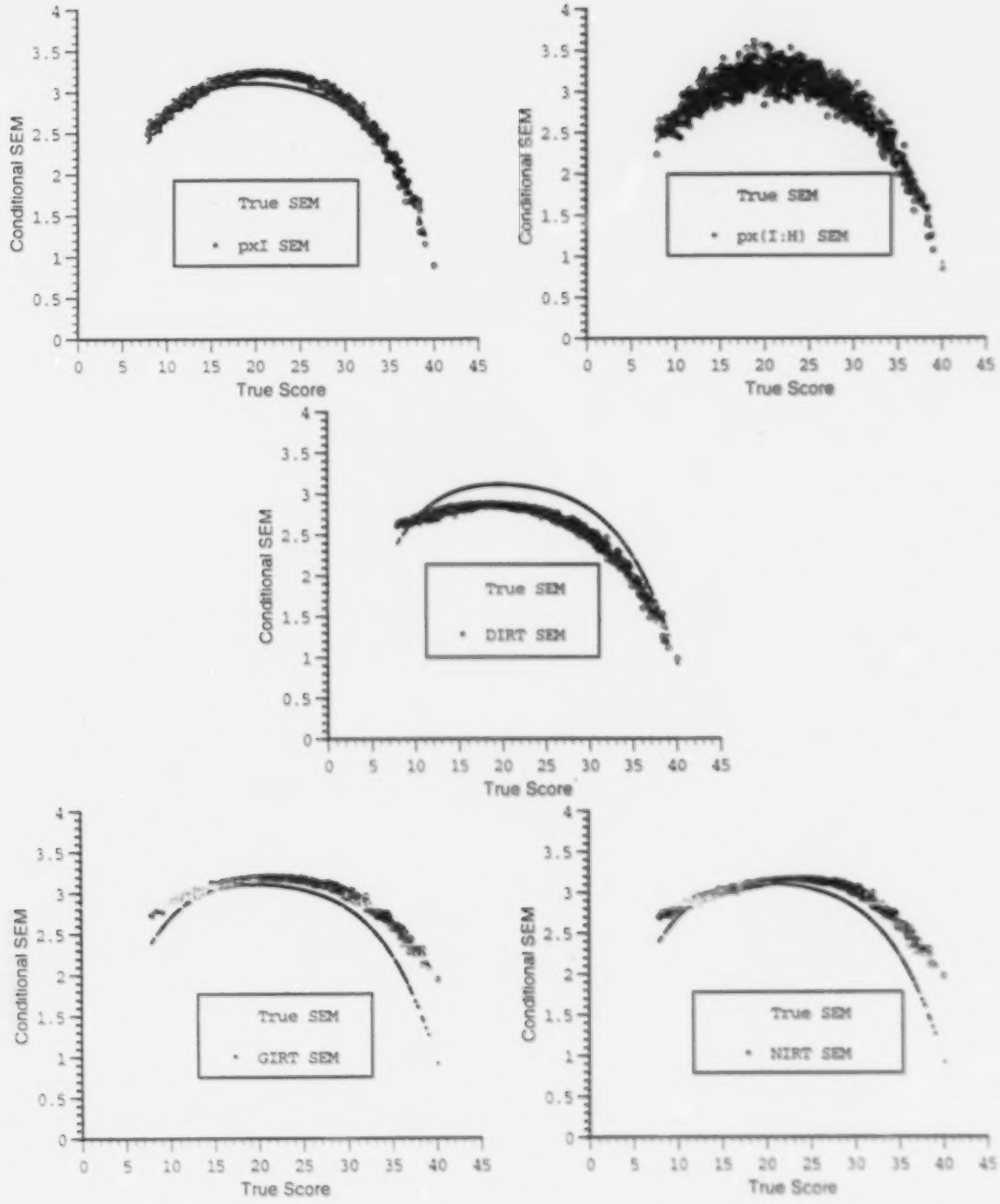

Figure 1. Comparisons of true conditional standard error of measurement and the mean of estimated conditional standard errors of measurement over 50 replications using five estimation methods and ksi=0.275.

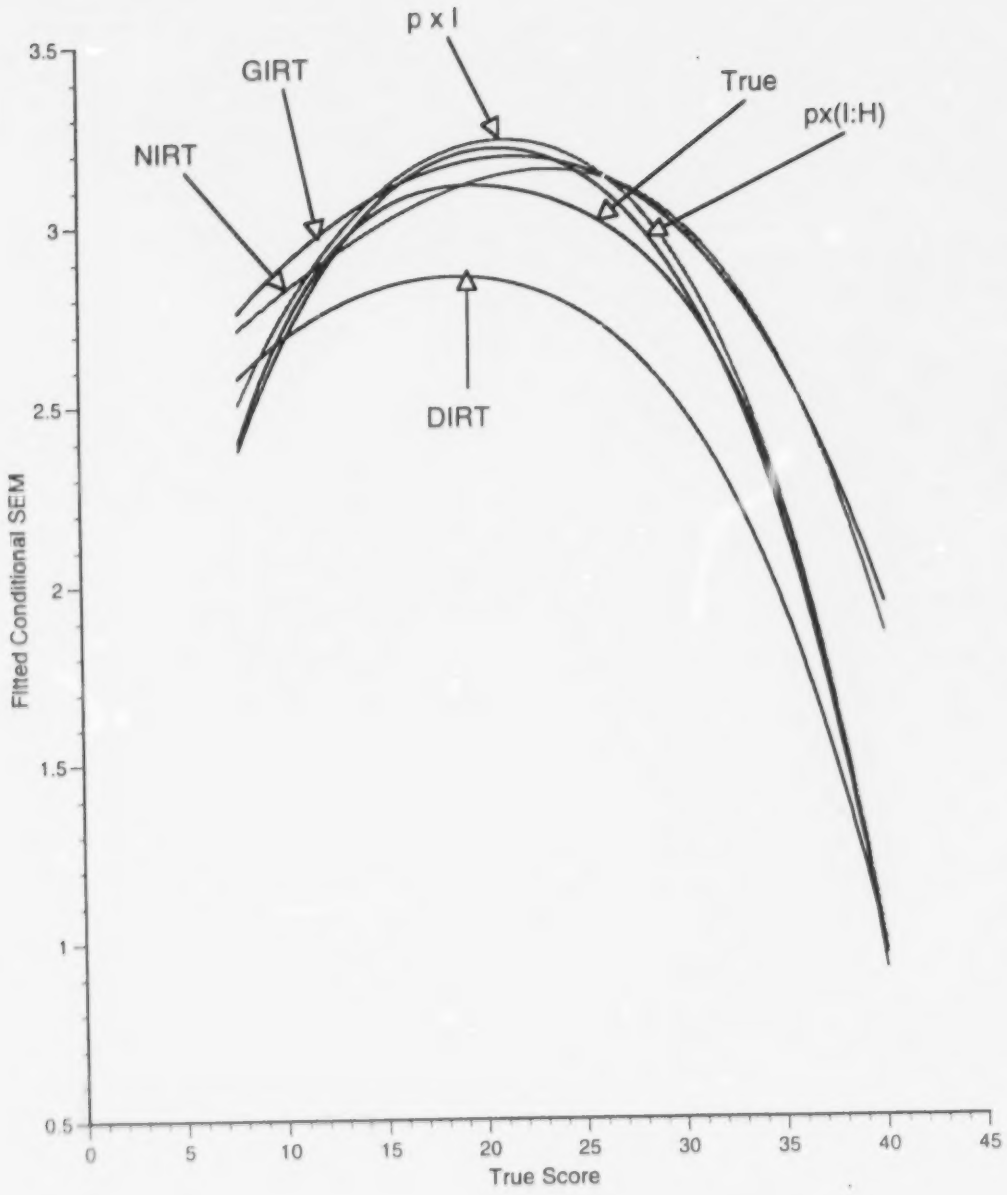

Figure 2. True conditional standard error of measurement and fitted conditional standard error of measurement for five estimation methods and ksi=0.275.

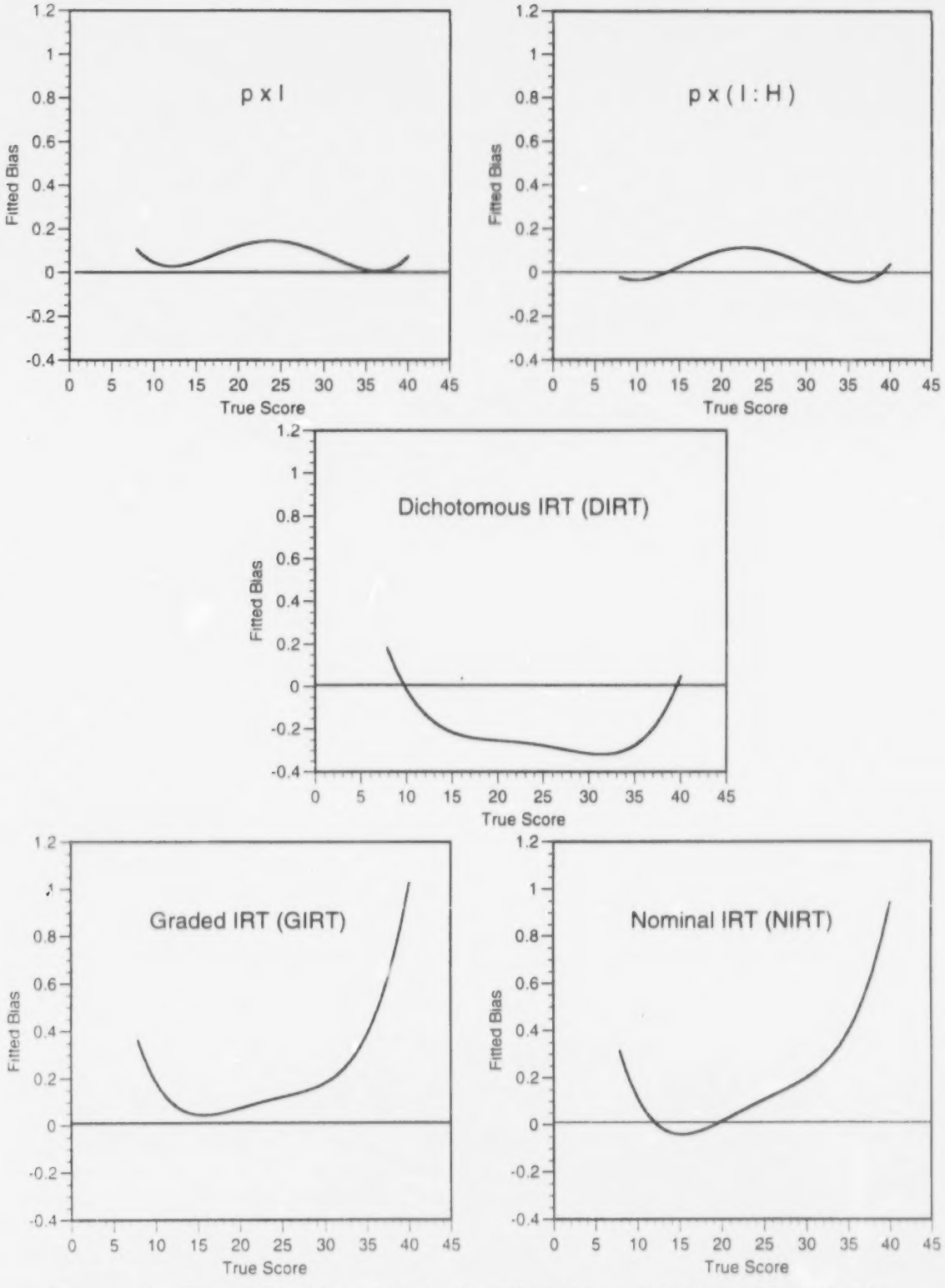

Figure 3. The bias line for each estimation method relative to the true conditional standard error of measurement for ksi=0.275.

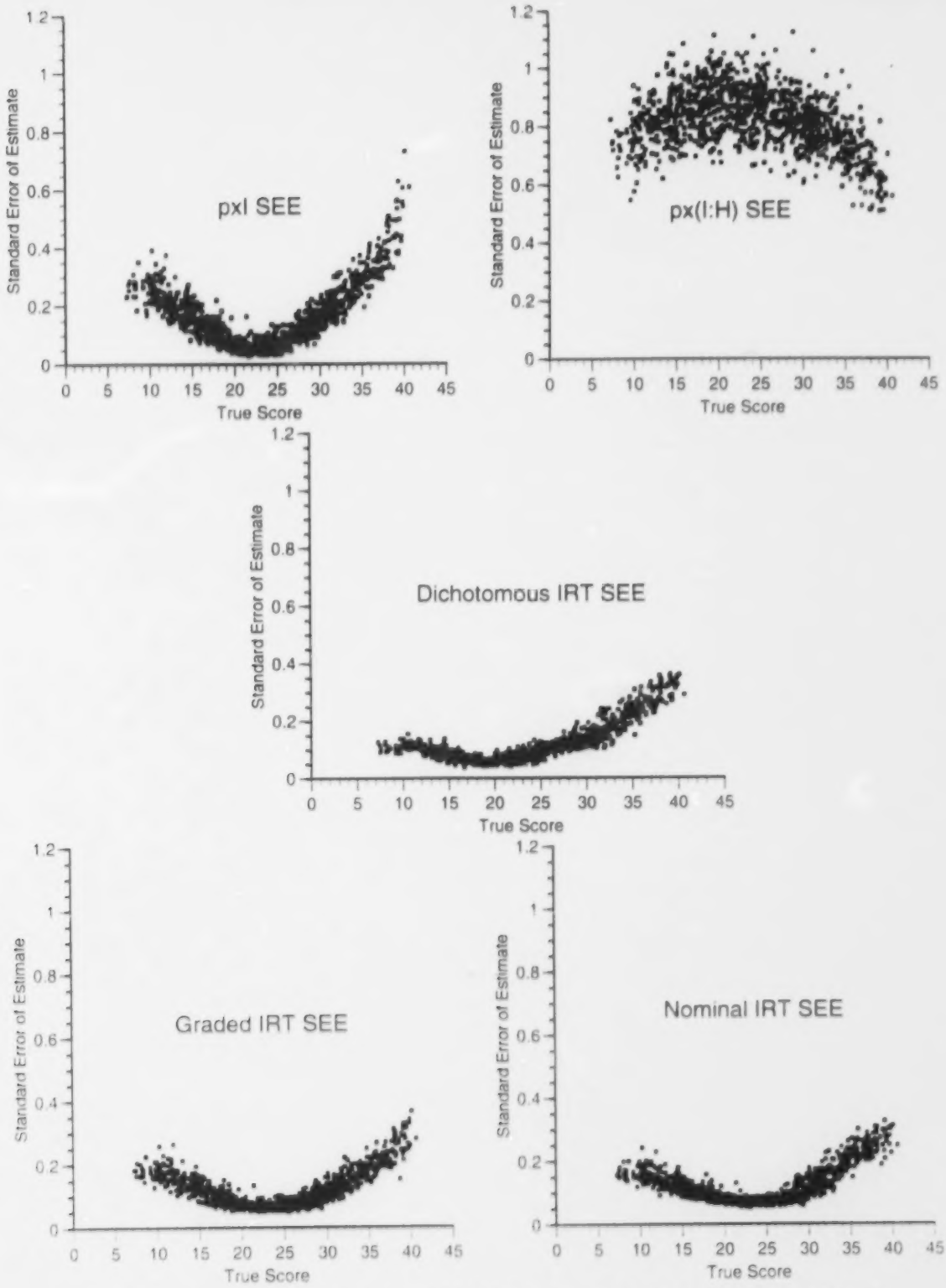

Figure 4. Standard error of estimate of each estimation method over 50 replications for ksi=0.275.

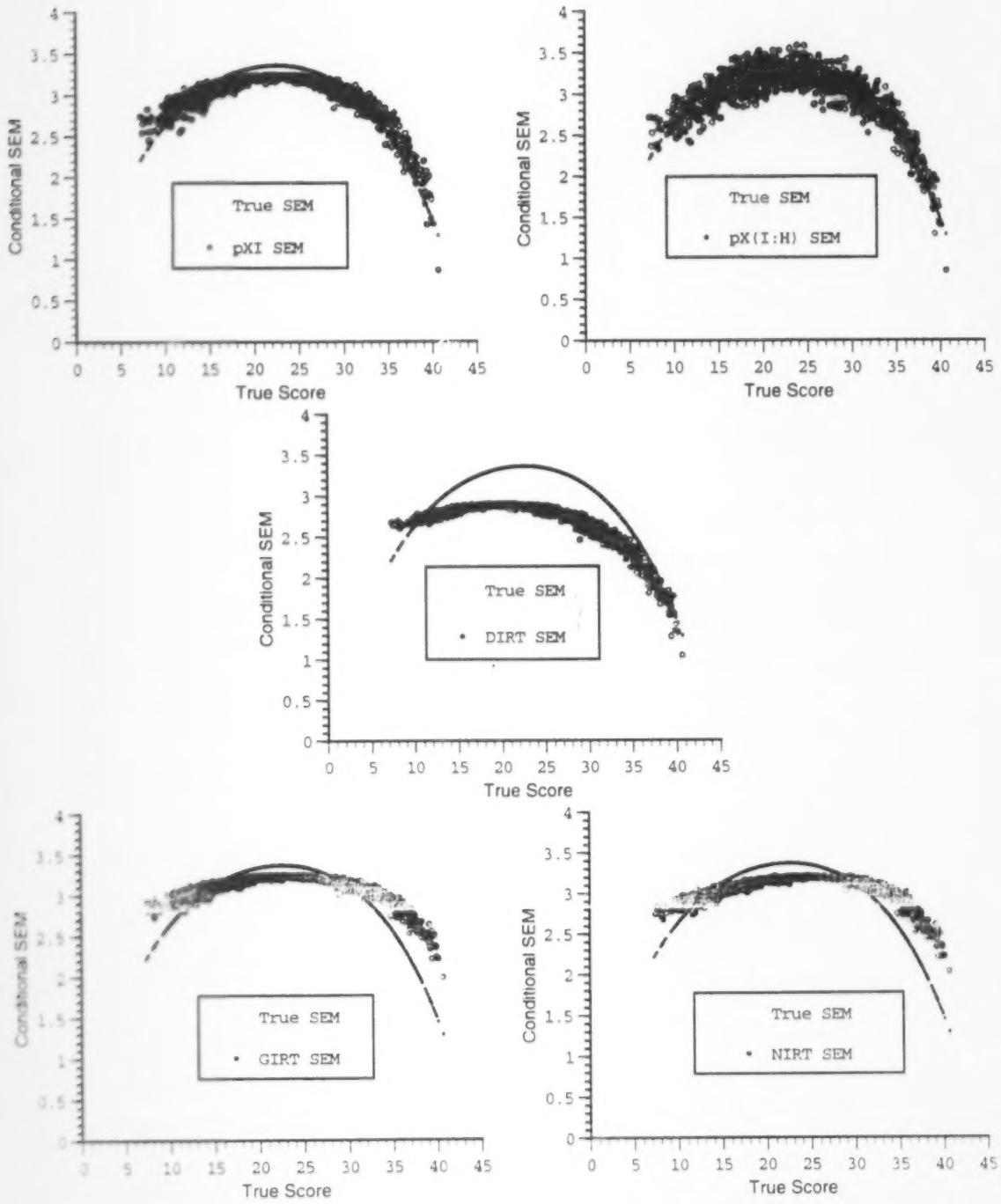

Figure 5. Comparisons of true condtional standard error of measurement and the mean of estimated conditional standard errors of measurement over 50 replications using five estimation methods and ksi=0.300.

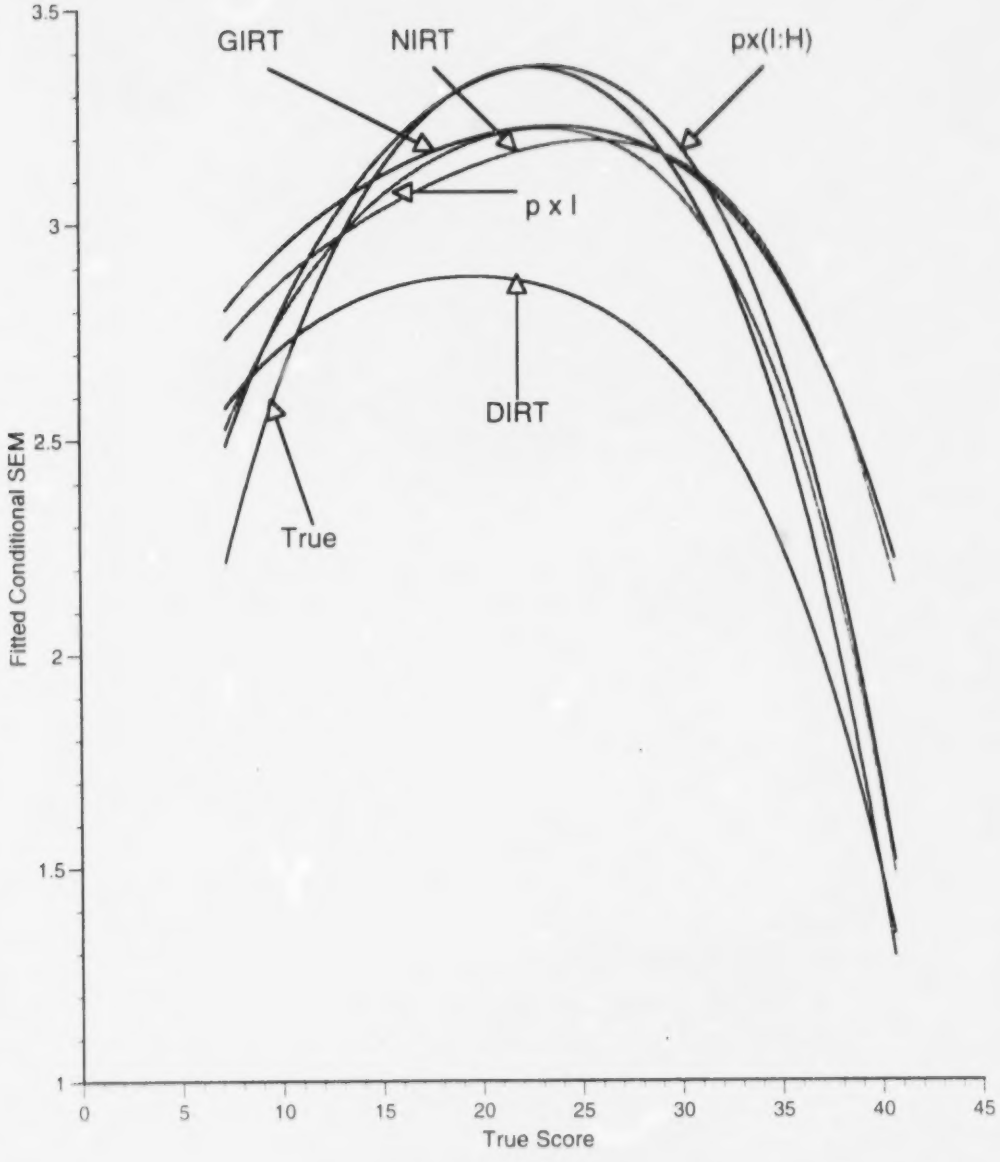

Figure 6. True conditional standard error of measurement and fitted conditional standard error of measurement for five estimation methods and ksi=0.300.

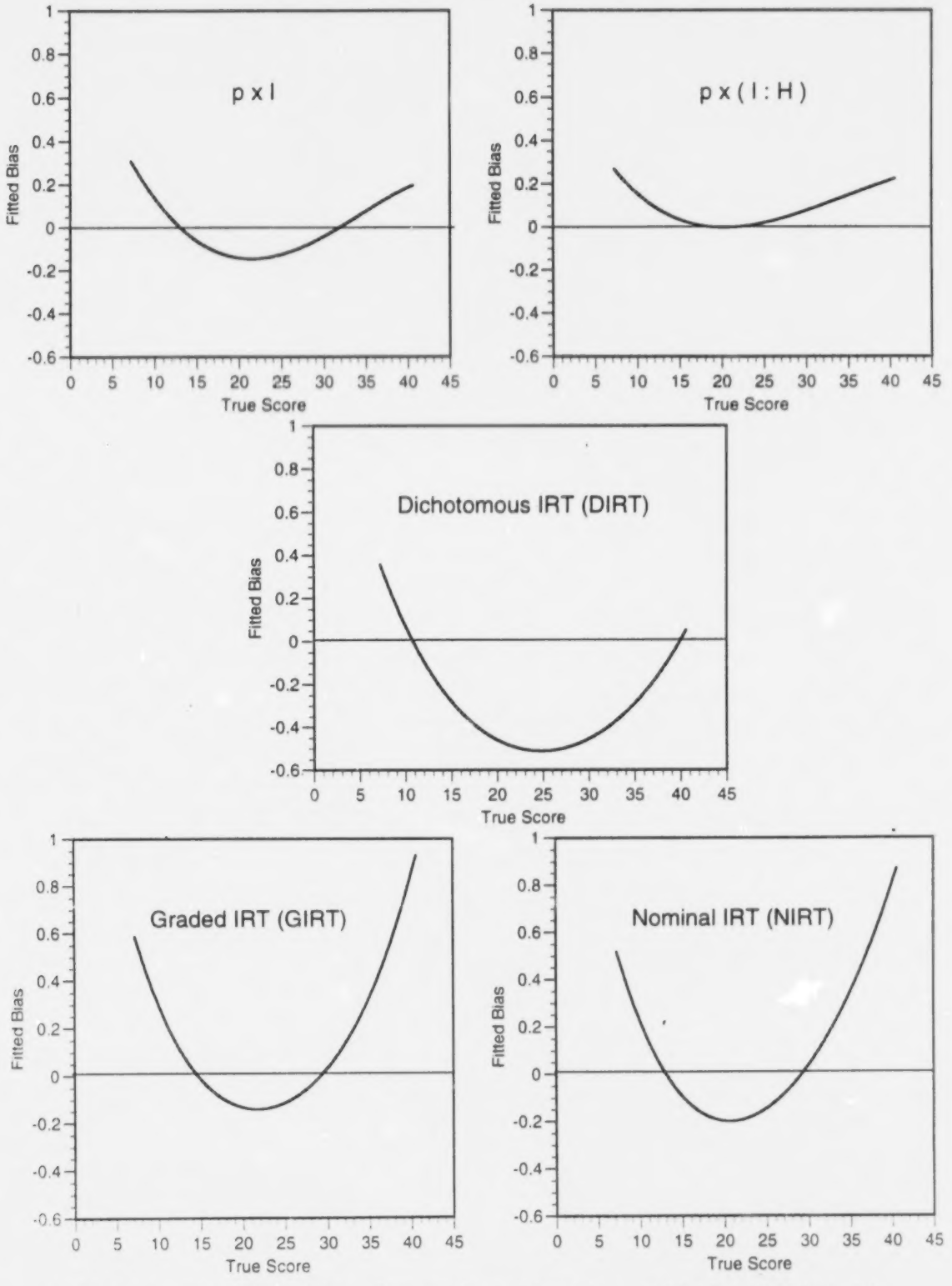

Figure 7. The bias line for each estimation method relative to the true conditional standard error of measurement for ksi=0.300.

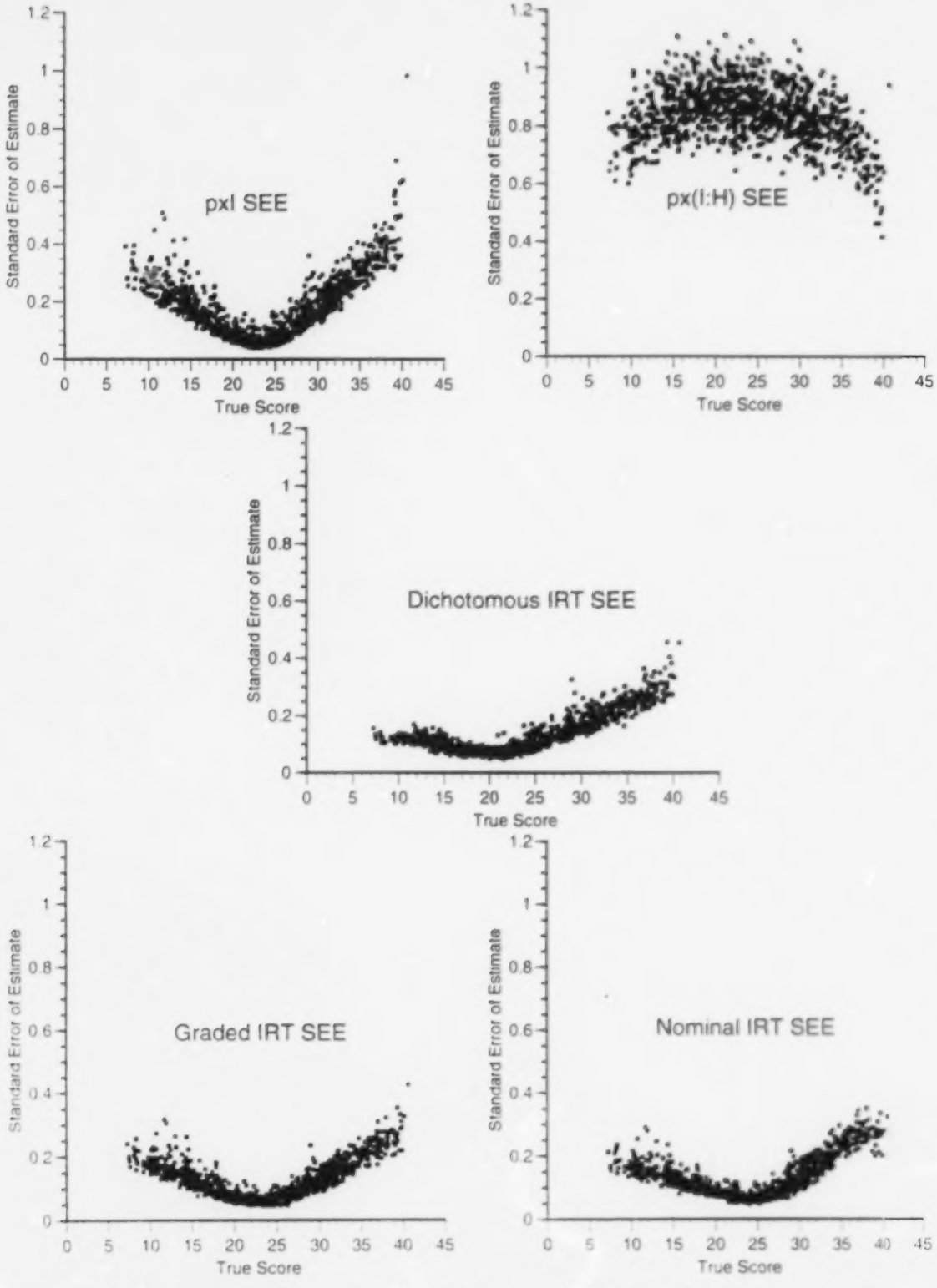

Figure 8. Standard error of estimate of each estimation method over 50 replications for ksi=0.300.

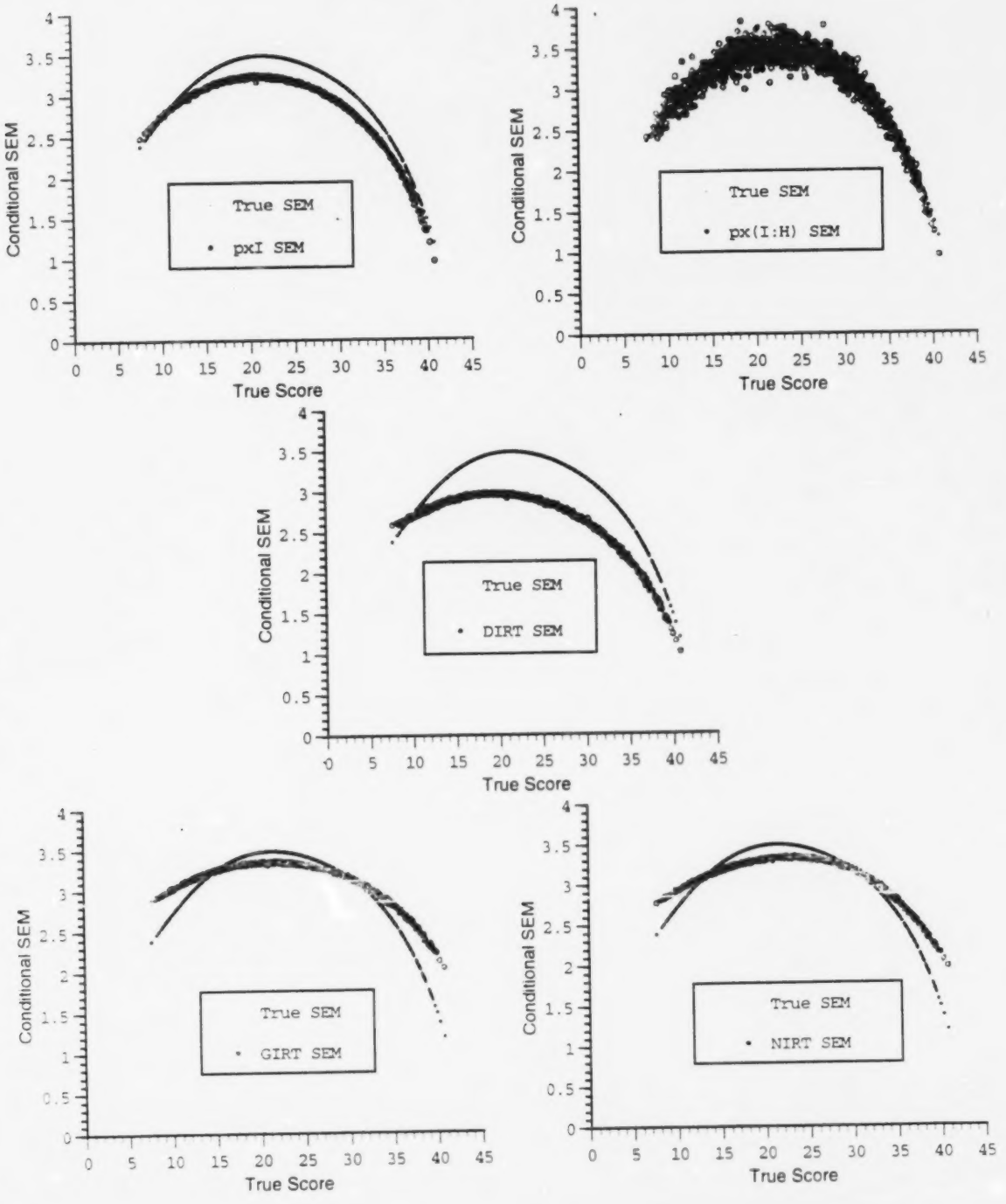

Figure 9. Comparisons of true conditional standarad error of measurement and the mean of estimated conditional standard errors of measurement over 50 replications using five estimation methods and ksi=0.325.

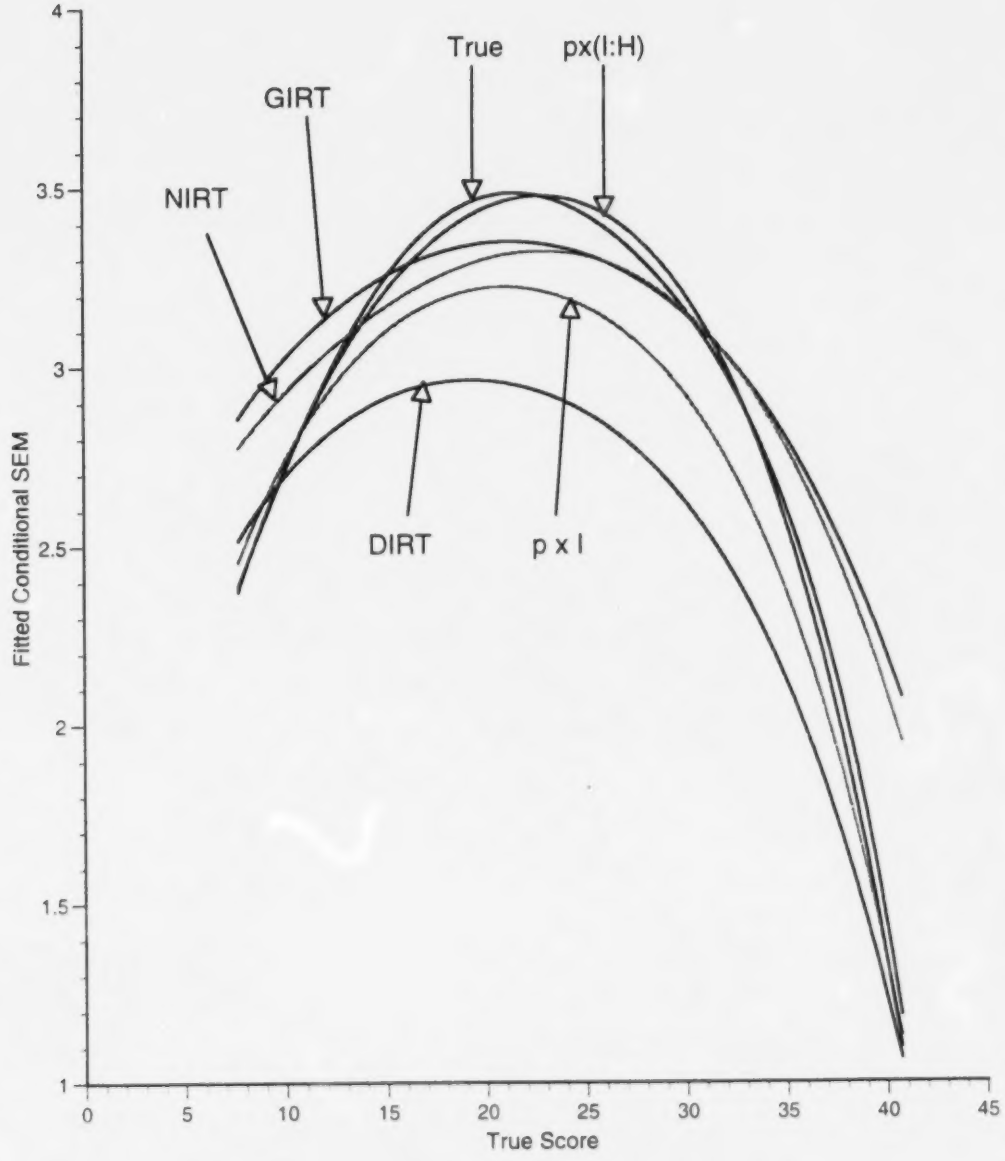

Figure 10. True conditional standard error of measurement and fitted conditional standard error of measurement for five estimation methods and ksi=0.325.

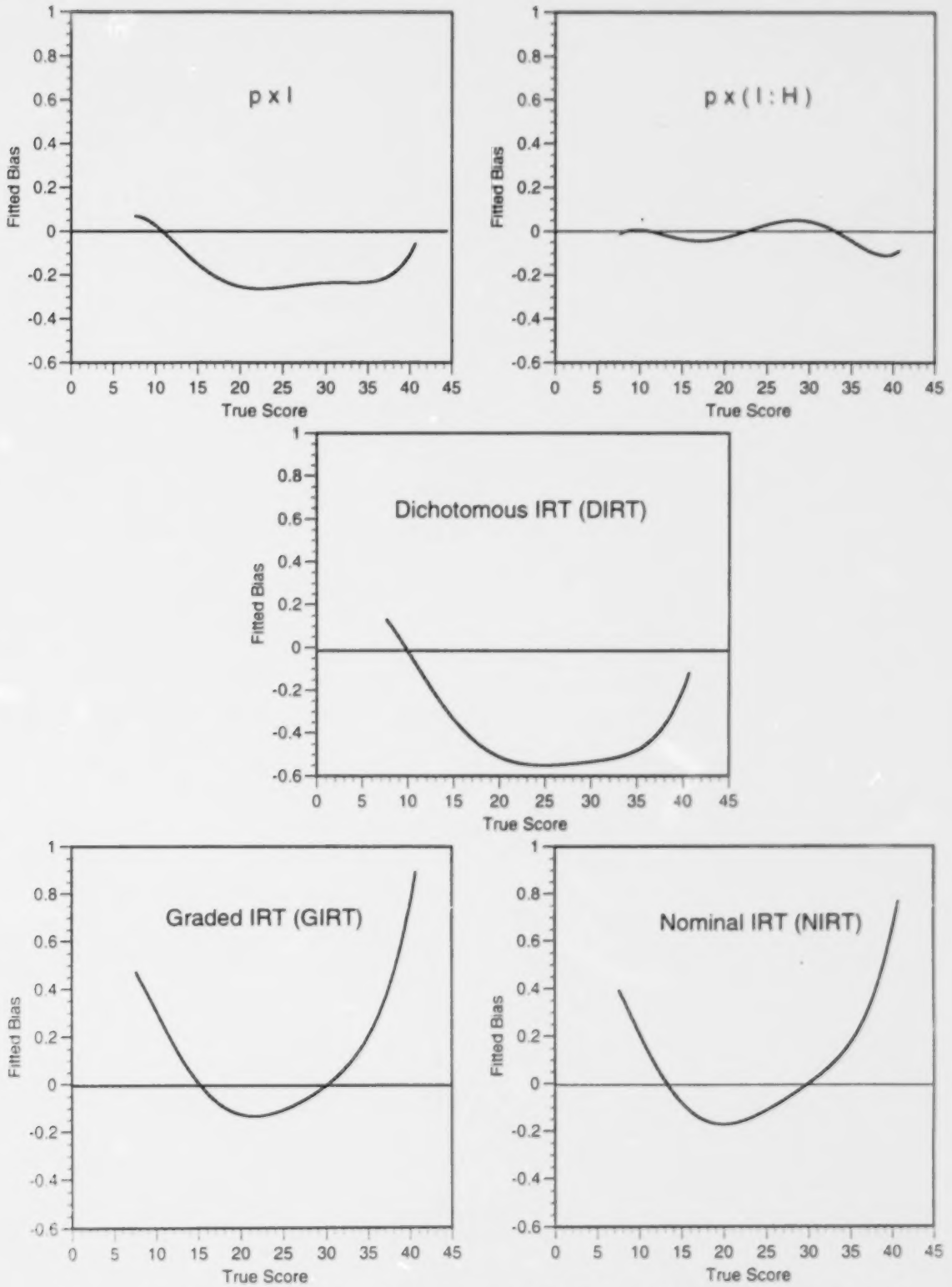

Figure 11. The bias line for each estimation method relative to the true conditional standard error of measurement for ksi=0.325.

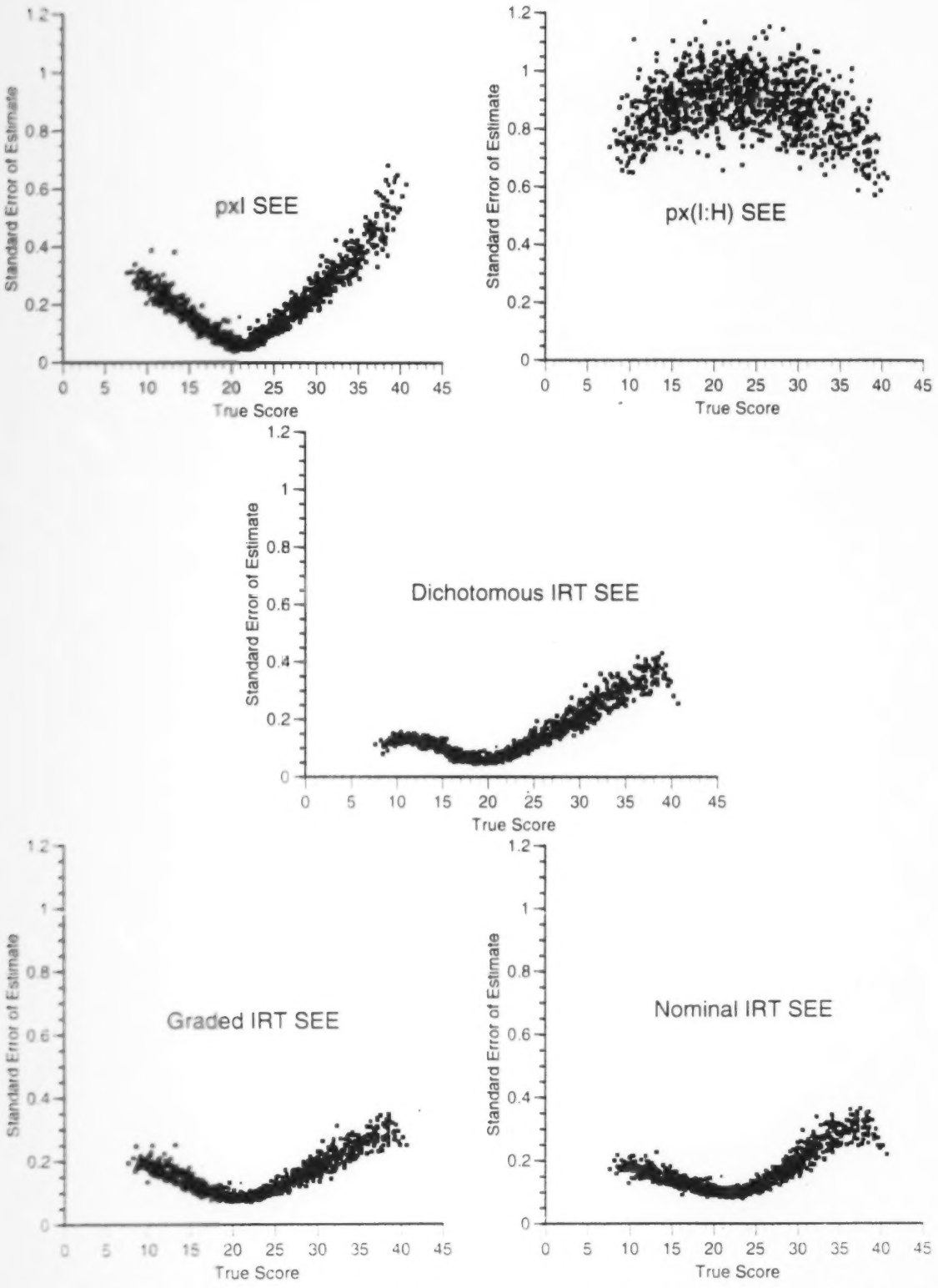

Figure 12. Standard error of estimate of each estimation method over 50 replications for ksi=0.325.

BEST COPY AVAILABLE

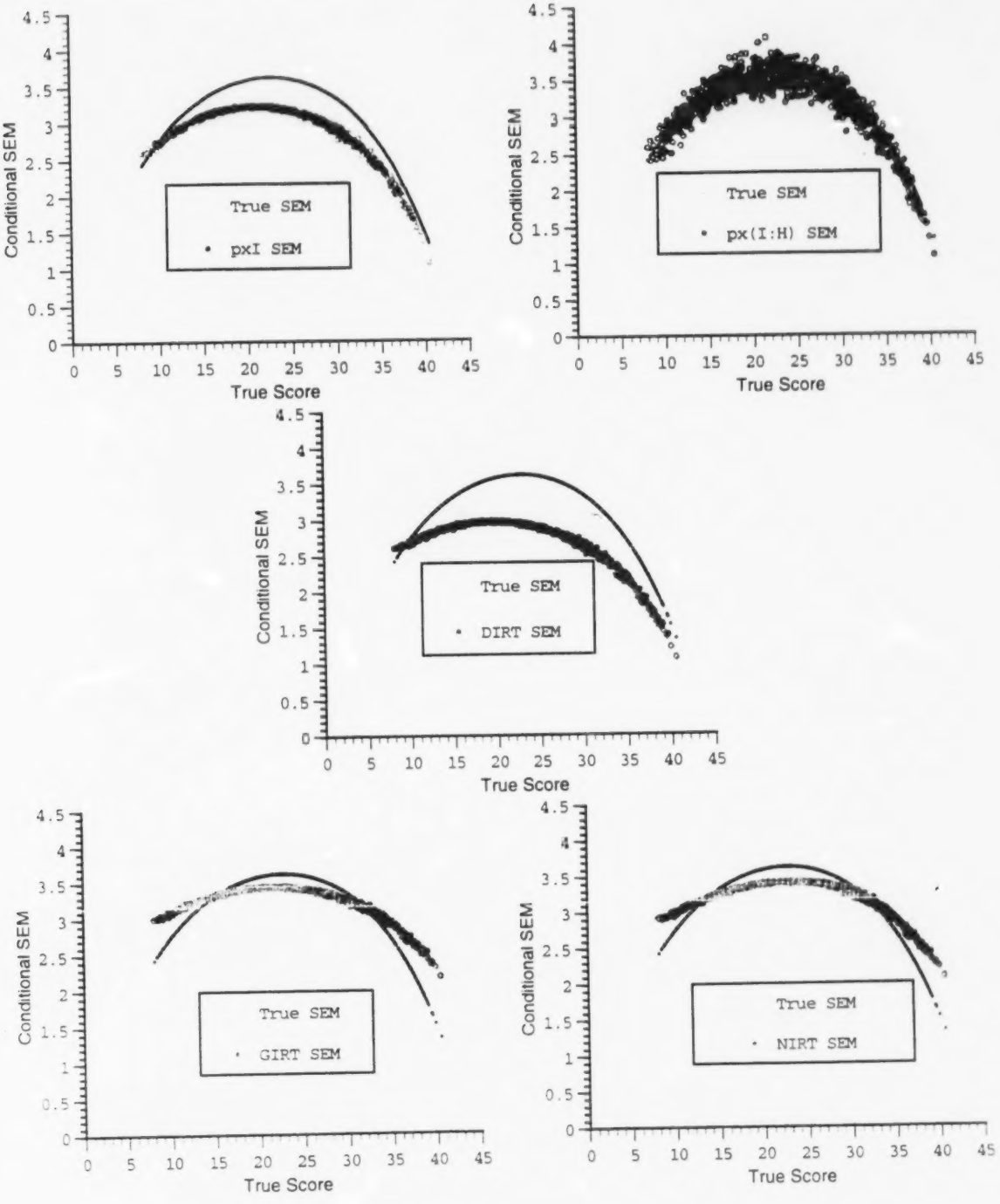

Figure 13. Comparisons of true conditional standard error of measurement and the mean of estimated conditional standard errors of measurement over 50 replications using five estimation methods and ksi=0.350.

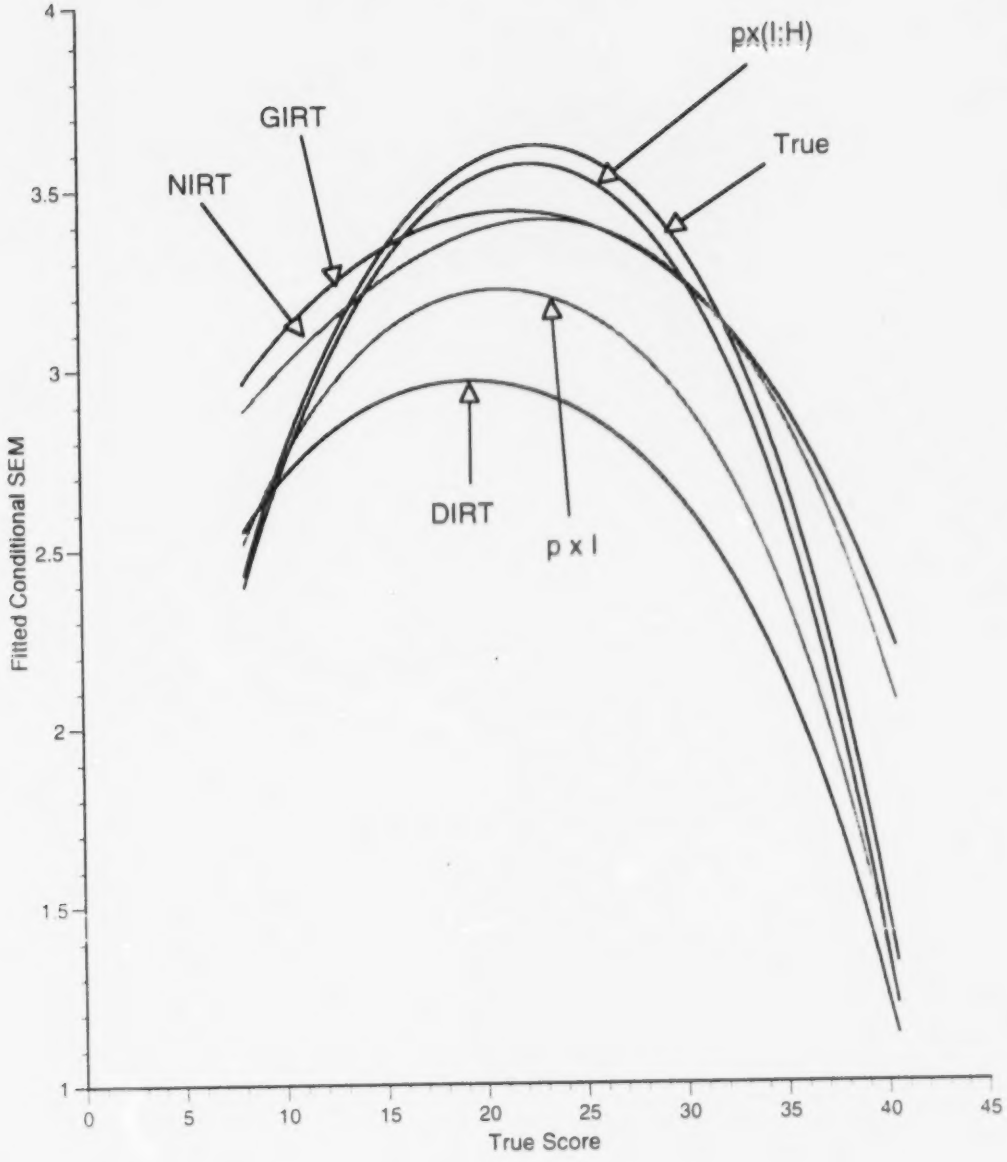

Figure 14. True conditional standard error of measurement and fitted conditional standard error of measurement for five estimation methods and ksi=0.350.

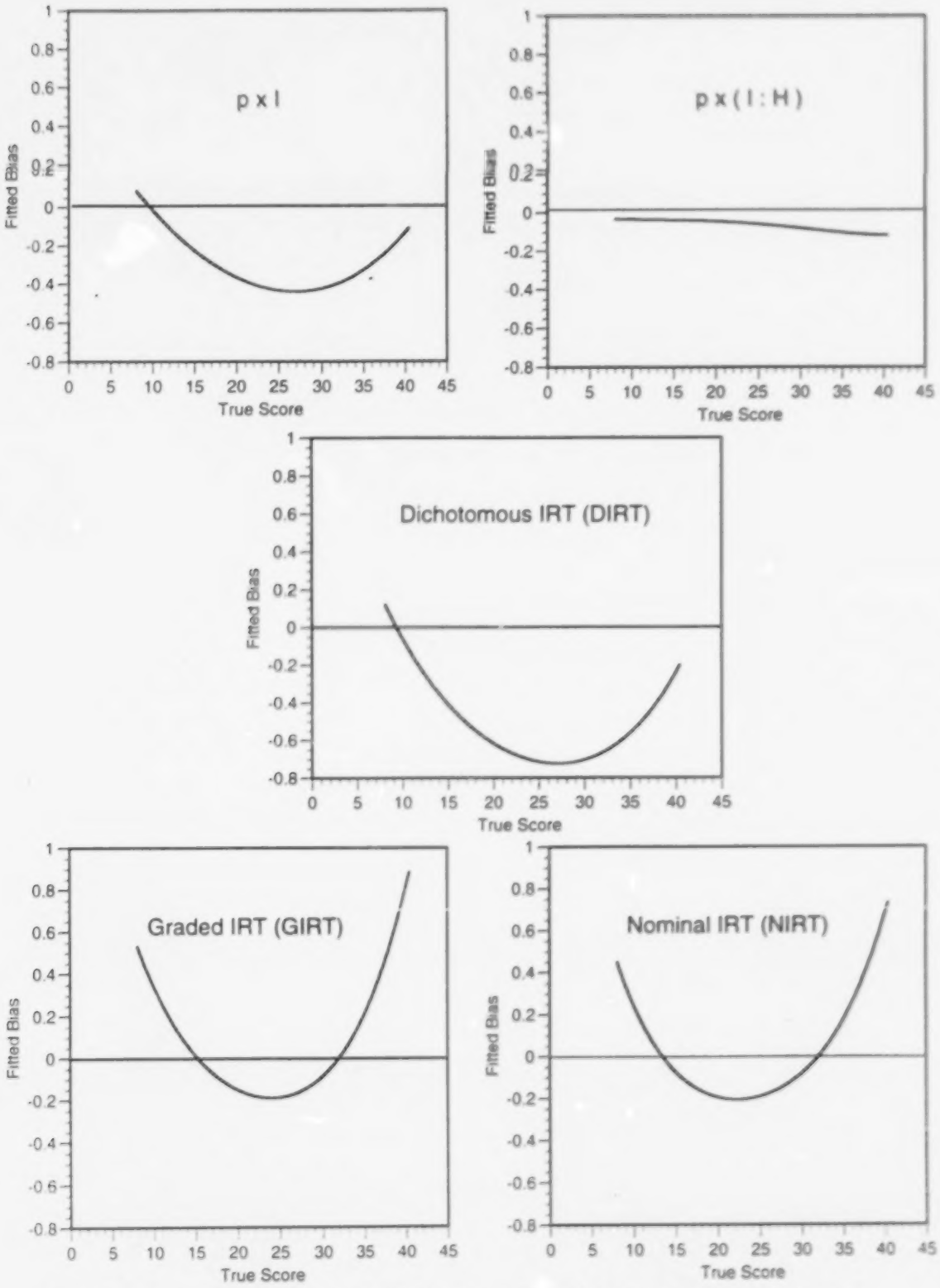

Figure 15. The bias line of each estimation method relative to the true conditional standard error of measurement for ksi=0.350.

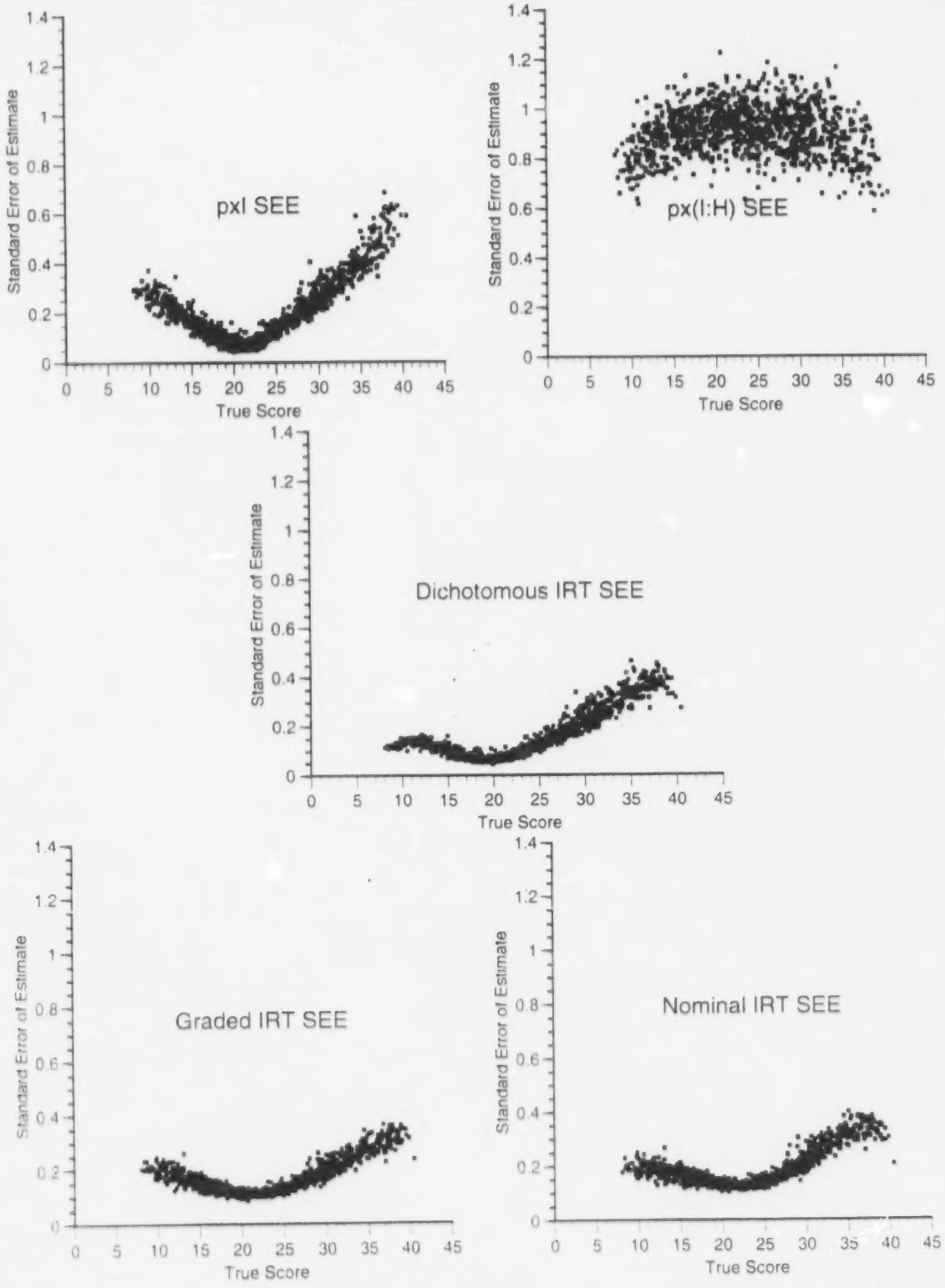

Figure 16. Standard error of estimate of each estimation method over 50 replications for ksi=0.350.

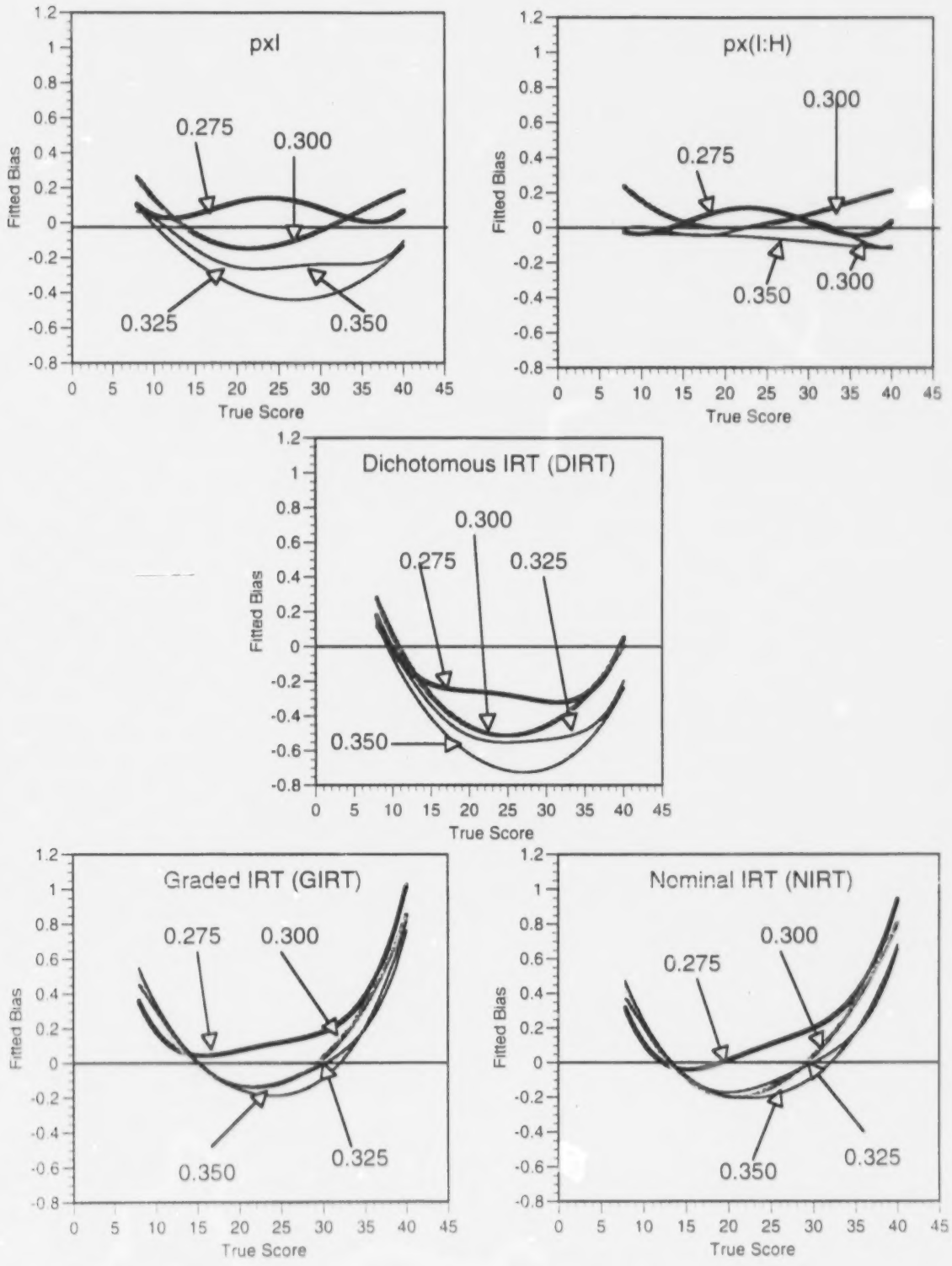

Figure 17. Comparison of bias lines for four specified ksi values within each of five estimation methods.

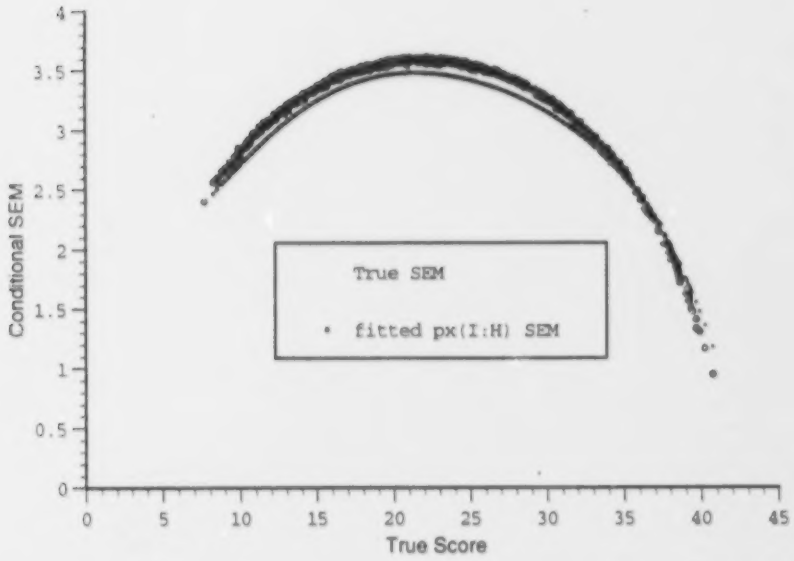

Figure 18. Comparison between the true conditional standarad error of measurement and the mean of estimated conditional standard errors of measurement for the fitted px(I:H) method.

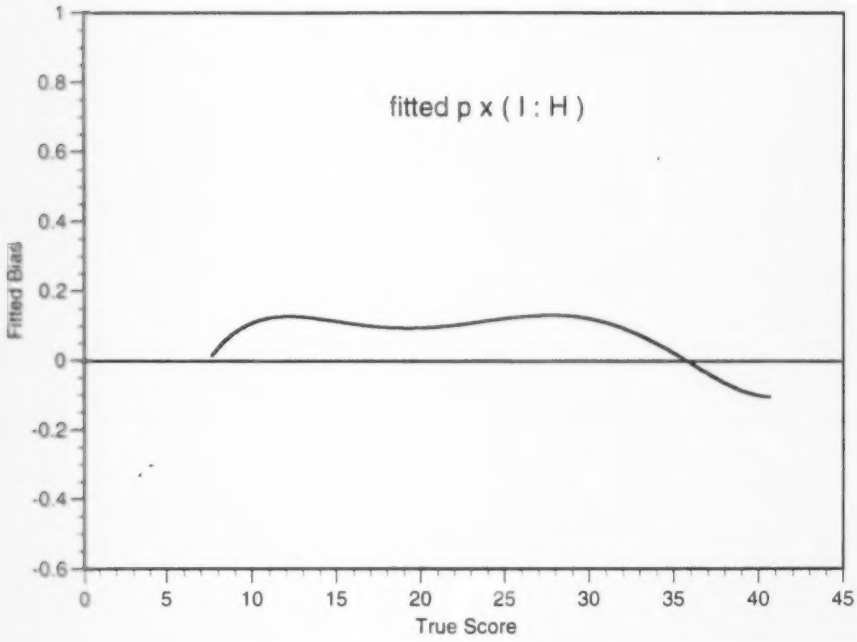

Figure 19. The bias line for the fitted px(I:H) estimation method compared to the true conditional standard error of measurement.

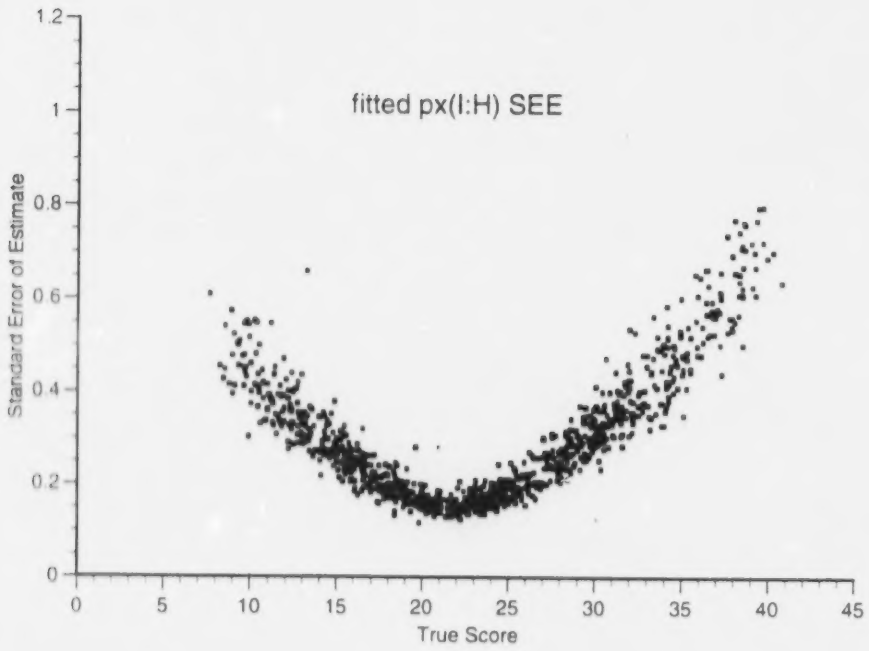

Figure 20. Standard error of estimate of the fitted px(I:H) method.

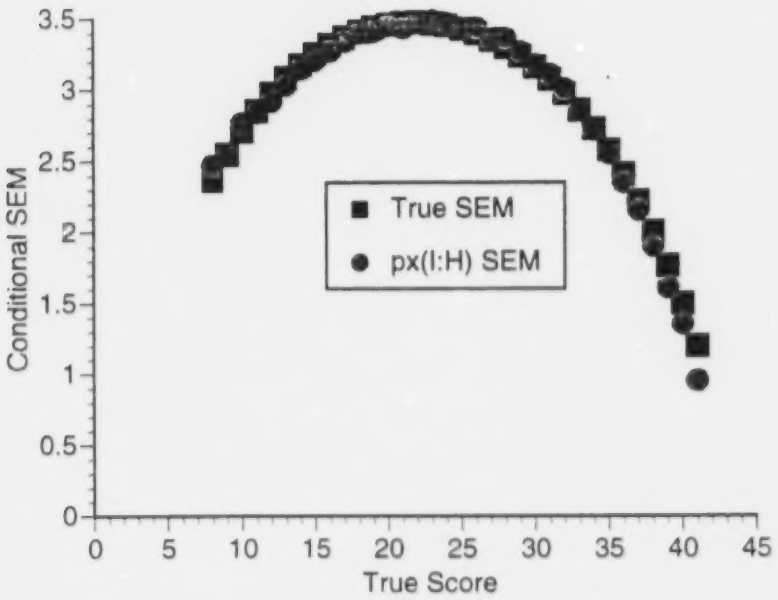

Figure 21. Comparison between the true condtional standard error of measurement and the mean of estimated conditional standard errors of measurement for the px(I:H) method using only integer score points.

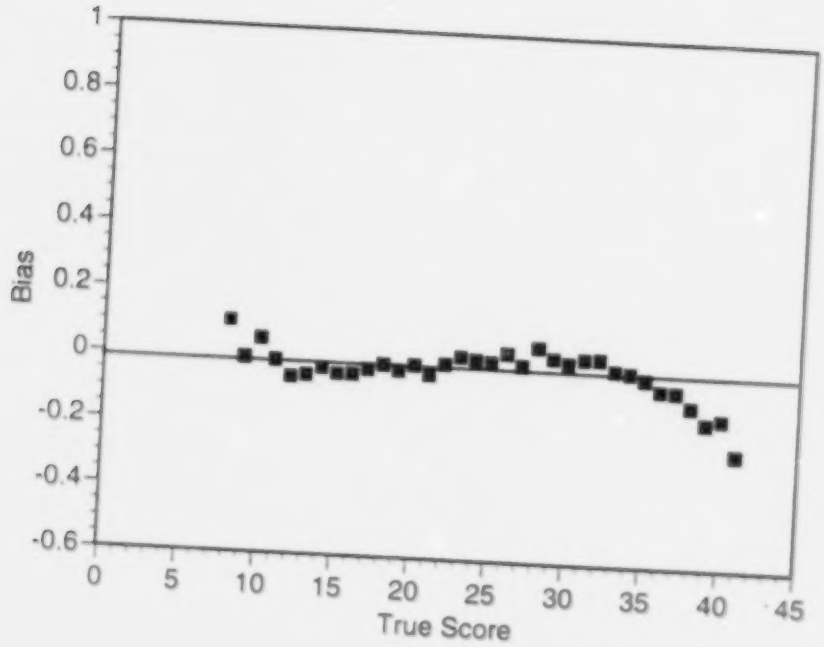

Figure 22. Bias of the px(I:H) estimation method compared to the true conditional standard error of measurement using only integer score points.

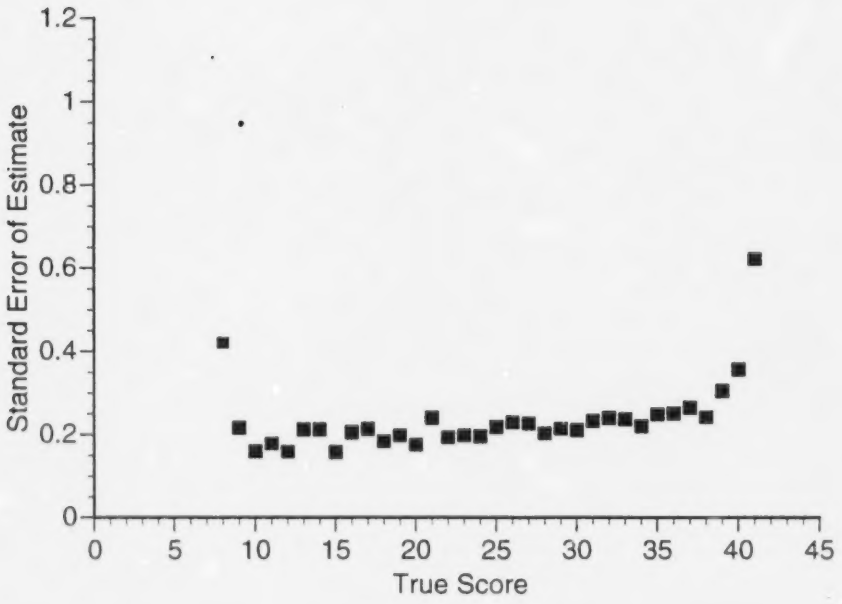

Figure 23. Standard error of estimate of the px(I:H) estimation method using only integer score points.

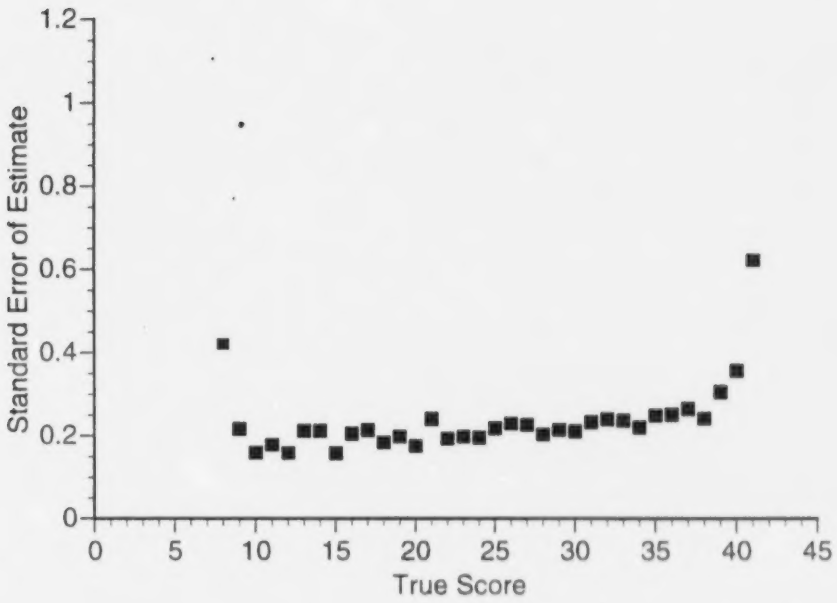

Figure 23. Standard error of estimate of the px(I:H) estimation method using only integer score points.

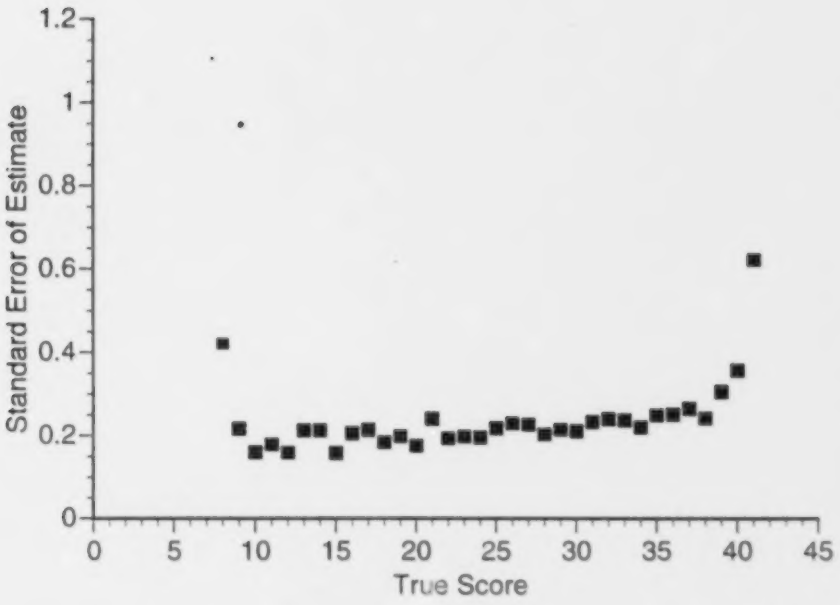

Figure 23. Standard error of estimate of the px(I:H) estimation method using only integer score points.

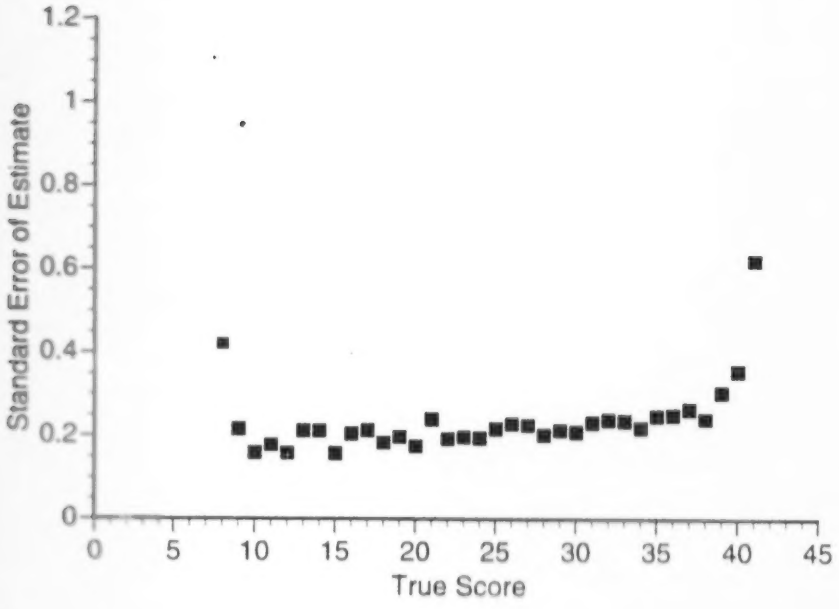

Figure 23. Standard error of estimate of the px(I:H) estimation method using only integer score points.

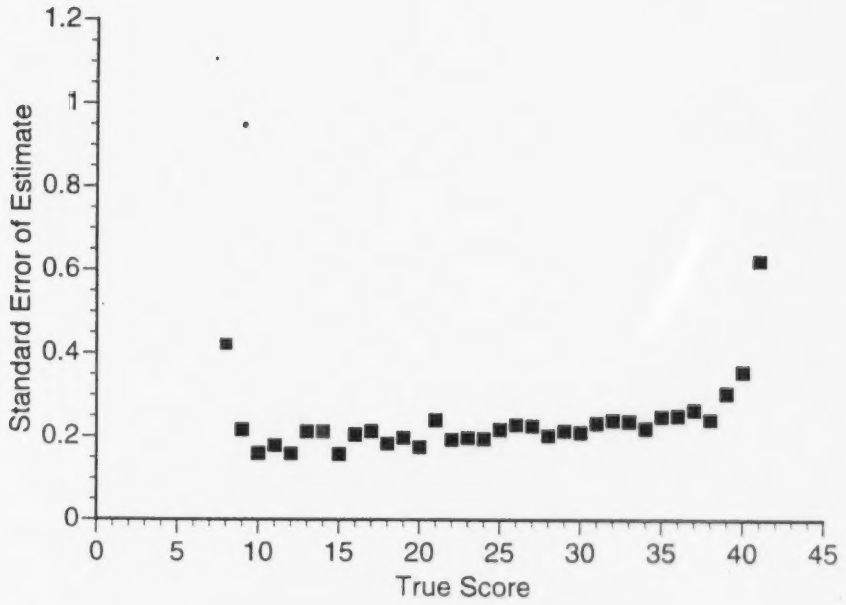

Figure 23. Standard error of estimate of the px(I:H) estimation method using only integer score points.